Département des systèmes agroalimentaires et ruraux CIRAD-SAR

RAPPORT DE STAGE
DEA de BIOSTATISTIQUE

ANALYSE DES DONNEES ET SYSTEMES D'INFORMATION GEOGRAPHIQUE

Maître de stage : M. PASSOUANT

JURY : M. G. GARAUX
Mme F. KAZI-AOUAL
M. P. MONESTIEZ
Mme G. VIGNAU
M. R. SABATIER

Stéphane THOMAS
ENSAM

CIRAD-SAR/N°45.94
Juillet 1994

Je tiens, tout d'abord, à remercier M. Michel Passouant, responsable du laboratoire informatique et biométrie du CIRAD-SAR, pour la confiance qu'il m'a témoignée, pour l'encadrement et le soutient qu'il m'a offert tout au long de ce stage.

Mme Christine Lavit et M. Robert Sabatier, responsables du DEA de biostatistique, m'ont été d'un grand secours en me guidant pendant mon stage, leurs aides m'ont été précieuses, je leur suis reconnaissant.

Je n'oublie pas Valérie Versini, secrétaire de l'unité de recherche CEEI, et Michel Bangil, assistant informatique du CIRAD-SAR, qui ont toujours été d'une très grande disponibilité et d'une patience à toute épreuve.

Enfin, je remercie toutes les personnes du CIRAD qui ont participées au bon climat qui a régné pendant la durée de ce stage. Je pense, en particulier, à François Deumier, Christine et Frédérique De Savoye, Isabelle fouquart, Olivier Constantin, Stéphane Degrés, et tous les autres.

Sommaire

Apports de l'analyse des données à l'étude de données cartographiables dans le cadre d'un Système d'Information Géographique

Un Système d'Information Géographique (S.I.G.) est un système informatique de gestion de données repérées dans l'espace, structuré de façon à pouvoir en extraire des synthèses utiles à la décision. Il y a donc cohabitation de données :

* géographiques (localisation des objets dans l'espace) sous forme de cartes,
* attributaires (description des objets repérés).

Dans son fonctionnement, un S.I.G. gère les opérations :

* d'acquisition des données,
* de gestion et de structuration des données,
* de traitement des données,
* de sortie et d'échange des données.

Le traitement statistique des données, dans le cadre des S.I.G., est actuellement assez limité : il est possible de réaliser des tris sur les variables, des calculs de moyennes et d'écarts types, et éventuellement de générer des graphes.

L'analyse des données est un outil très utile pour l'interprétation et la compréhension de phénomènes complexes. Nous allons voir, dans quelles limites, il est possible de l'adapter, dans un objectif exploratoire, aux contraintes des S.I.G. Tout d'abord, il est nécessaire de préciser la notion de données et d'objets, avant d'exposer les outils statistiques utilisables et leurs principes mathématiques. L'application d'une de ces méthodes à un cas particulier viendra compléter cette étude.

I. Objets manipulés par les S.I.G. et objectifs.

Afin de ne pas trop compliquer l'étude, nous considèrerons que les problèmes d'interprétation du réel (forme de la terre, projection, repérage dans l'espace, etc) ont été antérieurement résolus, et que les données géographiques sont sous la forme de cartes.

A. Les données géographiques.

1. Modes de gestion des données géographiques dans les S.I.G.

Actuellement, dans les S.I.G., deux modes de gestion des données géographiques cohabitent : le mode objet (structure vectorielle) et le mode image (structure raster). Ce fait est important, il conditionnera la façon d'exploiter les données géographiques par l'analyse des données.

Le mode objet (structure vectorielle) :

La carte est enregistrée sous la forme d'un plan structuré. Les données sont traitées telles que l'homme les conçoit : elles sont organisées en points, segments, arcs de cercles, surfaces, etc. Ces objets sont décrits par des points d'un plan affine, des équations et des conditions d'appartenance.

Le mode image (structure raster) :

La carte est découpée suivant une trame qui constitue un quadrillage régulier (en anglais : raster). A chacune des cases élémentaires ainsi définies est attribuée une valeur :

* 0 ou 1 : la case sera affichée en blanc ou en noir,
* 0 à n : n niveaux de gris,
* ou un code couleurs.

Cette valeur donne la dimension thématique : les cellules appartenant à un même objet reçoivent la même valeur.

Le format A4 (21 cm * 29.7 cm) avec une résolution de 300 dpi (300 points par pouce) représente 9 millions d'informations élémentaires. Les sorties de certaines méthodes importantes d'acquisition des données sont en mode raster : scanner, images satellites.

Un changement de mode de gestion des données est toujours possible, mais cette étape supplémentaire est coûteuse en temps et entraîne une déformation de l'information.

2. L'espace géographique.

Bien que dans l'absolu les trois dimensions de l'espace physique soient équivalentes, les S.I.G. actuels traitent l'altitude différemment des deux coordonnées horizontales. La raison de cette situation est double : le temps de calcul nécessaire au traitement de données en trois dimensions reste limitant, et les S.I.G. s'inscrivent dans la droite ligne de la cartographie plane.

Deux façons de gérer l'altitude sont couramment utilisées :

* Le Modèle Numérique de Terrain (M.N.T.) : sur la base d'un maillage régulier de l'espace, chaque noeud reçoit une valeur de l'altitude. Le pas du maillage peut aller de 50 à 100 m., les altitudes sont souvent déterminées par étude de photographies aériennes ou satellitaires.

* Le "Triangulated Irregular Network" (T.I.N.) : c'est un réseau irrégulier de triangles qui permet notamment de mémoriser les points importants du relief qui ne se trouvent généralement pas sur les noeuds du maillage M.N.T.

Une telle gestion de l'altitude n'est pas sans conséquences : ces deux modèles tendent à lisser le relief, et le calcul d'isolignes (courbes de niveau) nécessite la définition et l'utilisation d'une méthode d'interpolation. Pour un relief marqué, ces deux phénomènes conjugués ne semblent pas gênants, par contre, pour des régions de plaine, le relief est déformé et mal restitué.

Ne pouvant donc pas considérer les trois dimensions de l'espace comme entièrement équivalentes, nous ne traiterons, dans cette étude, que des données liées à des objets plans.

3. Interprétation géométrique des objets.

D'un point de vue purement géométrique, les objets qui sont représentés sur les cartes s'organisent en trois catégories : les points, les lignes, et les régions (figure n°1). Ces objets géométriques se retrouvent dans les multiples thèmes qui peuvent donner lieu à une cartographie : hydrographie, infrastructure (routes, chemins de fer, etc), cadastre, couverture végétale et cultures, géologie, pédologie, population, climat, etc. Suivant leur nature, les objets géométriques ne seront pas traités de la même manière dans l'analyse des données.

a. Les points.

Les points ne sont qu'une interprétation, une simplification du réel : tout objet réel a trois dimensions. Le point est utilisé pour représenter des objets réels qui, à l'échelle considérée, sont de dimensions négligeables par rapport aux autres objets. Un changement d'échelle peut entraîner un changement de représentation : un point peut devenir une surface. *Les points sont utilisés pour représenter, par exemple, les villes, les points de relevés pédologiques, les points d'eau, etc.*

b. Les lignes.

Les lignes sont utilisées pour symboliser des objets dont l'épaisseur est négligeable à l'échelle utilisée. Comme les points, les lignes peuvent devenir des régions lors d'un changement d'échelle.

Afin de représenter les formes physiques complexes (exemple : sinuosité d'une rivière), les lignes sont composées d'objets géométriques simples : segments de droites, arcs de cercles, parties de courbes polynomiales ou d'ellipses, etc.

Ces lignes peuvent être assemblées en *polylignes* pour des raisons de fonctionnalité : par exemple, une autoroute est une polyligne composée de lignes : les tronçons. Chacun des tronçons est plus ou moins sinueux, donc composé d'un nombre plus ou moins élevé d'objets géométriques simples.

Le type de polyligne le plus évolué est le *graphe* qui sert communément à représenter

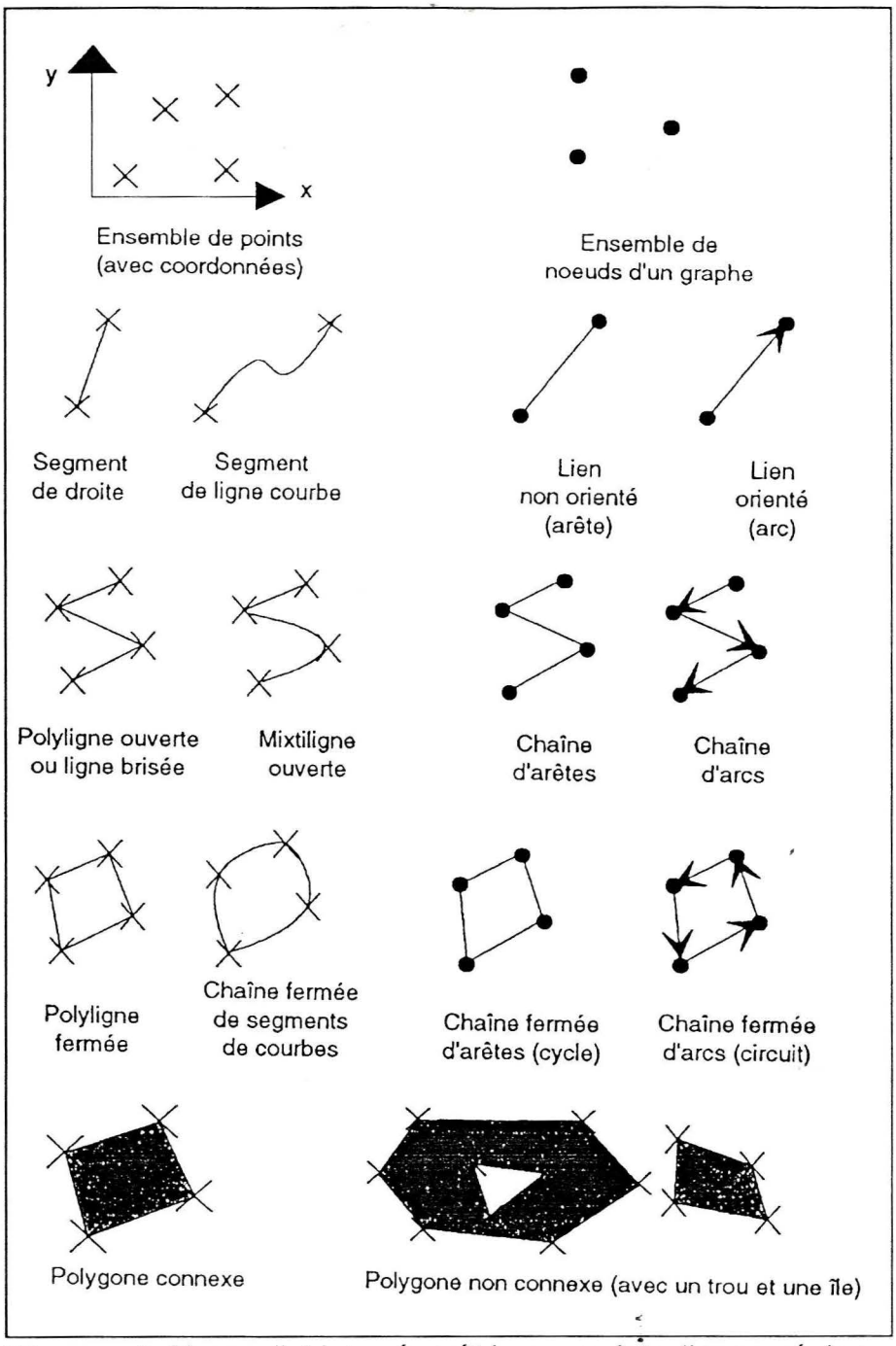


Figure 1: Types d'objets géométriques: points, lignes, régions.

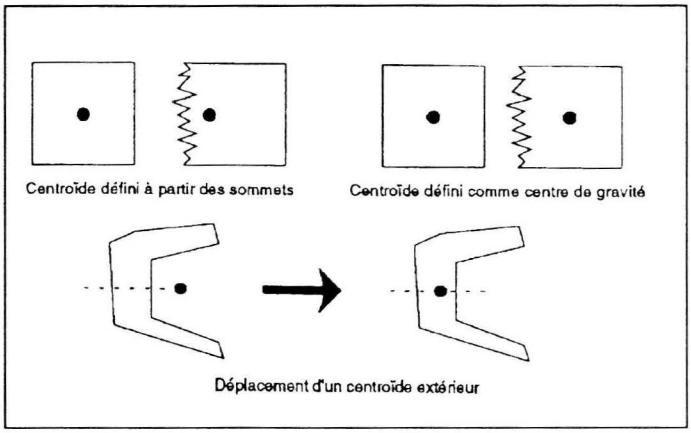


Figure 2: Types de centroïdes.

les réseaux (exemples : hydrographie, routes, lignes E.D.F., etc). Les graphes peuvent être orientés (hydrographie) ou non orientés (réseau téléphonique). Dans un graphe, la terminologie utilisée est différente (figure n°1) :

* les points sont des noeuds,
* les segments non orientés sont des liens,
* les segments orientés sont des arcs.

Certains graphes peuvent être cycliques ou acycliques.

Les caractéristiques géométriques importantes des lignes sont leurs longueurs, leurs orientations (lignes orientées), et leurs inter-relations (assemblage en graphes, en polylignes). *Les lignes symbolisent des objets physiques aussi divers que les rivières, les routes, les réseaux électriques, etc, mais aussi des objets immatériels tels que les champs de potentiel (exemple : champ de vent).*

Il faut noter qu'il y a divergence entre les lignes que nous venons de définir, et les lignes au sens informatique graphique. Ces dernières sont en fait ce que nous avons appelé objets géométriques simples : segments, arcs de cercles, etc.

c. Les régions.

Les régions symbolisent des objets qui ont deux dimensions importantes à l'échelle considérée.

Elles sont délimitées par des lignes qui peuvent être des polylignes. Par exemple, les frontières d'un pays sont découpées en tronçons en fonction des différents voisins (pour la France : frontière avec l'Espagne, l'Italie, etc). Sur chacun de ces tronçons, un certain nombre d'informations sont connues. les frontières ne peuvent donc pas être traitées d'un seul bloc, en une ligne simple.

Une région se caractérise par son contour, sa surface, par un point particulier appelé centroïde (figure n°2), et par sa périphérie (notion de voisinage).

Le centroïde peut être défini, soit comme centre de gravité de la région, soit comme centre de gravité des sommets. S'il est à l'extérieur de la région qu'il caractérise (par exemple les polygones non connexes, ou non convexes), les géographes préfèrent alors le déplacer pour corriger cet état. Cette modification semble discutable, car la restriction d'une région à un point est, de toute façon, une déformation de l'information.

4. Relations entre objets géographiques.

a. Distances géographiques entre objets.

Nous appellerons *distance géographique*, la notion commune de distance, que nous relèverons sur les cartes.

Par définition, une *distance mathématique* vérifie les propriétés suivantes :

pour trois individus a,b,c,

1) $d(a,b) = d(b,a) \geq 0$,

2) $d(a,b) = 0 \Leftrightarrow a = b$,
3) $d(a,b) \leq d(a,c) + d(c,b)$ inégalité triangulaire.

Une distance géographique peut ne pas être une distance au sens mathématique : si l'on étudie le réseau routier à l'intérieur d'une ville, le trajet entre deux points A et B peut être différent d'un sens à l'autre en raison de sens interdits. Dans cette situation, $d(A,B) \neq d(B,A)$ et pourtant, c'est cette distance géographique qu'il faut prendre en compte dans une étude des relations entre A et B.

Les objets géométriques que nous avons étudiés précédemment, étant définis dans un espace affine à trois dimensions, l'expression :

$$d^2(a,b) = (x_a - x_b)^2 + (y_a - y_b)^2 + (z_a - z_b)^2,$$

définit une distance euclidienne. L'espace euclidien affine est le cadre de nombreuses méthodes de l'analyse des données.

Dans la réalité que retranscrit l'espace géographique, la distance euclidienne qui correspond à la distance "à vol d'oiseau", ne semble pas toujours adaptée.
Les exemples suivants montrent cette inadéquation :

* figure n°3 : distance entre deux villes par la route,
* figure n°4 : distance sur un réseau (électrique, téléphonique, informatique, etc),
* figure n°5 : distance entre deux villes par le rail (elle s'apparente à la distance mathématique dite "à centre").

De plus, la distance entre deux types d'objets géométriques dépend de la réalité physique des objets représentés. Si l'on étudie la distance entre un point et une région, on a les situations suivantes :

* distance maximum : c'est la distance que prend en compte une personne désirant irriguer une parcelle à partir d'un point d'eau,
* distance minimum : c'est la distance entre une ville et un lac.

La distance entre deux objets physiques peut être, elle aussi, envisagée de plusieurs façons différentes. Par exemple, la distance entre deux villes peut être :

* le temps minimum de trajet,
* la distance minimum,
* la distance par la route,
* la distance par le chemin de fer,
* la distance par l'air,
* la distance euclidienne.

La distance "temps minimum de trajet" est une notion difficile à gérer. Par exemple, dans le cas du réseau routier, elle n'est pas constante au cours du temps : les conditions de circulation varient (densité du trafique, conditions climatiques, etc). Elle n'est pas non plus constante sur un même type de route (nationale, départementale, etc), elle varie en fonction de la sinuosité. Cette notion peut être approchée par le décalage entre la distance "à vol

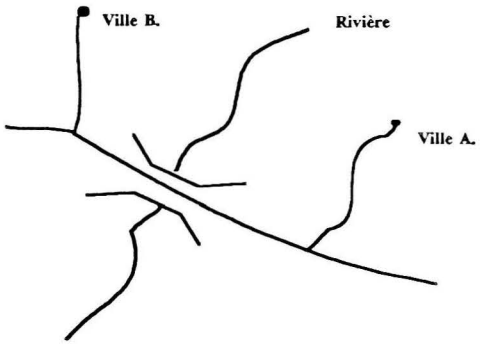


Figure n°3.

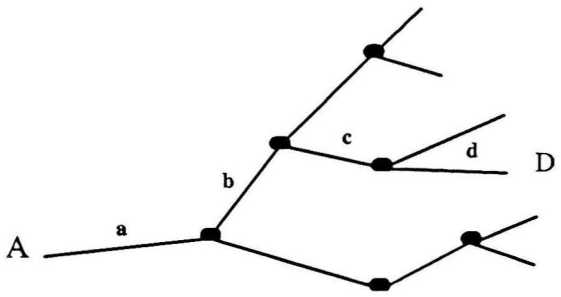


Figure n°4.

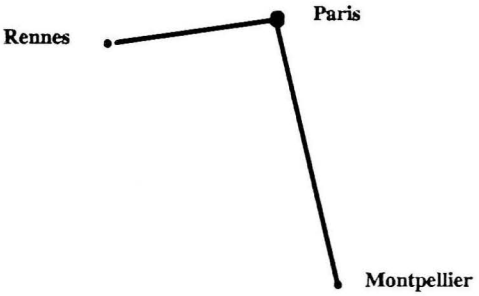
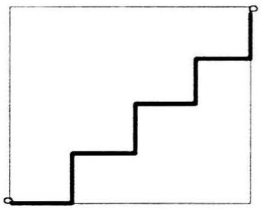


Figure n°5.

Forte sinuosité Faible sinuosité

Figure n°6



Distance de Manhattan

Exemples de distances géographiques

d'oiseau" et la longueur de la ligne (figure n°6). La distance "temps minimum de trajet" pose un autre problème que les géographes appellent "discontinuité ou déformation de l'espace" (figure n°7) : l'espace situé à moins d'une heure de transport de Paris forme un cercle autour de cette ville, mais l'espace situé à moins de deux heures de Paris forme un cercle concentrique autour de la ville, mais aussi un cercle autour de Lyon en raison de l'existence d'une ligne TGV sans arrêt entre les deux villes. Cela signifie que l'attraction créée par Paris se reporte autour de Lyon avant de se reporter autour de villes plus proches de Paris, mais moins bien desservies.

Il n'est donc pas possible de définir une distance-type entre les objets géométriques. Il est nécessaire de choisir les distances en fonction des thèmes à traiter et des objectifs poursuivis par l'étude. A ce stade, on ne doit pas baser les choix sur la distinction distance géométrique - distance mathématique.

b. Relations topologiques entre objets.

Toute l'information contenue sur une carte géographique ne peut pas être résumée par la seule notion de distance. Les positions relatives des objets entre eux représentent une notion importante.

Ces relations topologiques peuvent être standardisées en fonction des types géométriques d'objets. Il nous faut donc définir et relever sur la carte les relations suivantes :

Soient deux objets (points, lignes, ou régions) A et B,

L'inclusion : tous les points de A appartiennent à B : A est inclus dans B,

L'intersection : certains points de A appartiennent à B : A et B sont sécants,

Le voisinage : A et B ont une frontière commune. Deux régions seront dites voisines si leurs contours ont un point ou une ligne en commun, deux lignes seront voisines si elles ont une extrémité commune. Il est possible d'étendre cette notion aux graphes : deux noeuds seront voisins si ils sont reliés par un arc ou une arête. Le voisinage, ainsi défini, correspond à la notion mathématique de contiguïté.

Les deux dernières relations topologiques peuvent être quantifiées :

* l'intersection peut se mesurer en longueur de ligne incluse ou en surface commune,
* le voisinage peut se mesurer, dans la cas de régions, en longueur de frontière commune.

B. Les variables attributaires.

Les variables attributaires forment le deuxième type de données gérées par les S.I.G. Elles servent à définir la réalité thématique des objets. Ces variables sont quantitatives ou qualitatives.

1. Les variables attributaires attachées aux objets.

Lors de l'étude d'une ville les variables attributaires seront :

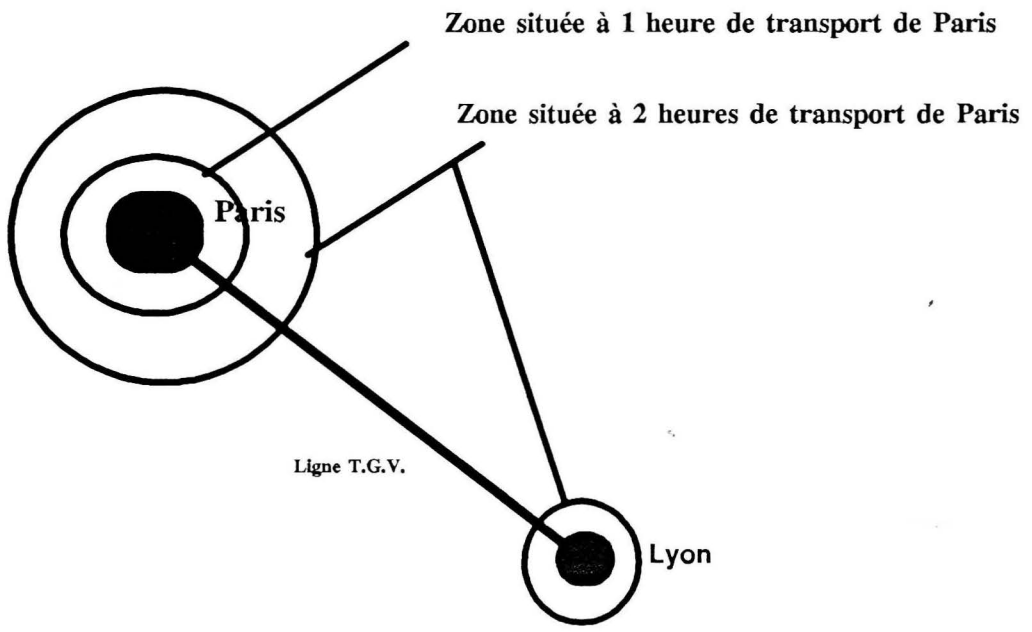


Figure n°7: discontinuité de la distance temps / trajet

* des variables simples : populations, infrastructures, capacités, etc,

* des tableaux : [population, revenus, nombre d'enfants, etc] croisés avec les catégories socio-professionnelles,

* des courbes : évolution de la population au cours du temps.

Donc, du point de vue mathématique, à un objet peuvent être attachés :

* les valeurs de plusieurs variables (un vecteur),

* des tableaux : individus - variables, ou variables - temps,

* des courbes. Elles seront toujours, ou pourront toujours être discrétisées et traitées comme des tableaux.

Le choix des variables attributaires utilisées pour décrire les objets physiques devra être réalisé en fonction des objectifs poursuivis par l'étude et des thèmes abordés.

Les objets géométriques sont, par nature, structurés (de petites régions s'assemblent pour former une grande région, des lignes composent une polyligne, etc), il faut noter qu'il peut exister, de la même manière, une structure sur les variables attributaires.

Il faut se méfier des biais introduits par la façon de mesurer les variables. Trois problèmes importants sont, plus particulièrement, liés aux variables attributaires dans les S.I.G. Les données sont très souvent incomplètes, il n'y a pas toujours eu de relevés réguliers. De nombreuses variables existent pour décrire le même phénomène (variables fortement corrélées), et ce ne sont pas toujours les mêmes qui ont été utilisées d'un objet physique à un autre. De plus, certaines variables peuvent ne pas être codées toujours de la même manière (exemple : le découpage en classes d'une variable quantitative peut avoir été fait de plusieurs façons différentes d'un objet à l'autre). Il serait donc nécessaire de standardiser la définition d'objets par des variables attributaires.

2. Variables attributaires attachées aux relations entre objets.

Un certain nombre de variables attributaires ne sont pas liées aux objets physiques, mais aux relations entre objets. Il s'agit notamment de tous les flux entre objets. Ces relations particulières sont matérialisées par les notions de voisinage et d'appartenance aux mêmes graphes.

Pour une autoroute, chaque tronçon reçoit une partie du trafic des tronçons voisins. Il n'y a pas indépendance entre les portions d'autoroute, c'est ce qu'il faut prendre en compte. Une relation entre tronçons sera caractérisée par :

* des variables simples : flux moyen de véhicules, de marchandises, de personnes, etc,

* des tableaux : [flux de véhicules, flux de marchandises, flux de personnes, etc] croisés avec les catégories de véhicules,

* des courbes : évolution d'un flux en fonction du temps.

Donc, les variables attributaires qui peuvent être attachées à une relation entre objets sont du même type que celles attachées aux objets.

C. Changements d'échelle et divisions de l'espace.

Ces deux opérations sont fréquemment réalisées dans les S.I.G., et ont des implications similaires au niveau des relation objets - données attributaires.

1. Les changements d'échelle.

Les changements d'échelle peuvent entraîner des changements de représentation des objets physiques : comme nous l'avons vu précédemment, les lignes ou les points peuvent devenir des régions et réciproquement.

Même si la représentation géométrique des objets change, ce sont les variables attributaires qui semblent plus difficiles à gérer. Par exemple, si l'échelle augmente, un ensemble de cantons deviendra une entité : le département. Il faut, en parallèle, agréger les variables attributaires : dans un cas simple, la somme de la population des cantons donnera la population du département, mais si la variable considérée est le salaire moyen par catégories socio-professionnelles (C.S.P.), il faut connaître les effectifs par C.S.P. dans les différents cantons (information supplémentaire), pour agréger la variable. Le problème se complique quand l'échelle diminue : il faut désagréger les variables. Dans ce cas, il faut définir des règles de désagrégation qui peuvent être complexes : il n'y a aucune raison pour que la population et le salaire moyen par C.S.P. soient répartis de façon homogène sur tous les cantons.

2. Les divisions de l'espace.

Une division de l'espace est généralement réalisée lors d'une superposition de plusieurs cartes thématiques. En effet, les découpages de l'espace géographique étant différents d'une carte thématique à l'autre, les objets ne se superposent pas et une nouvelle division plus complexe de l'espace est ainsi créée. Dans l'exemple de la figure 8, la fusion des informations des cartes géologique et agronomique impose la création de nouveaux objets : (parcelle A, sol n°1), (parcelle A, sol n°2), (parcelle B, sol n°1), (parcelle B, sol n°2). Le problème de désagrégation se pose de nouveau : le rendement de la culture réalisée sur A est le fait de deux sols distincts (sol n°1 et sol n°2), et la répartition entre les deux n'est pas connue.

D. Evolution des données au cours du temps.

1. Les données géographiques.

De nombreuses modifications de l'information géographique peuvent intervenir au cours du temps :

* apparition ou disparition d'objets (exemples : création d'axes de communication, abandon de parcellaires de conquête),

* déplacement d'objets (exemple : modification naturelle du lit d'une rivière),

* modification du mode de représentation (exemple : la représentation d'une ville, suite à sa croissance, peut passer de l'objet géométrique "point", à l'objet géométrique

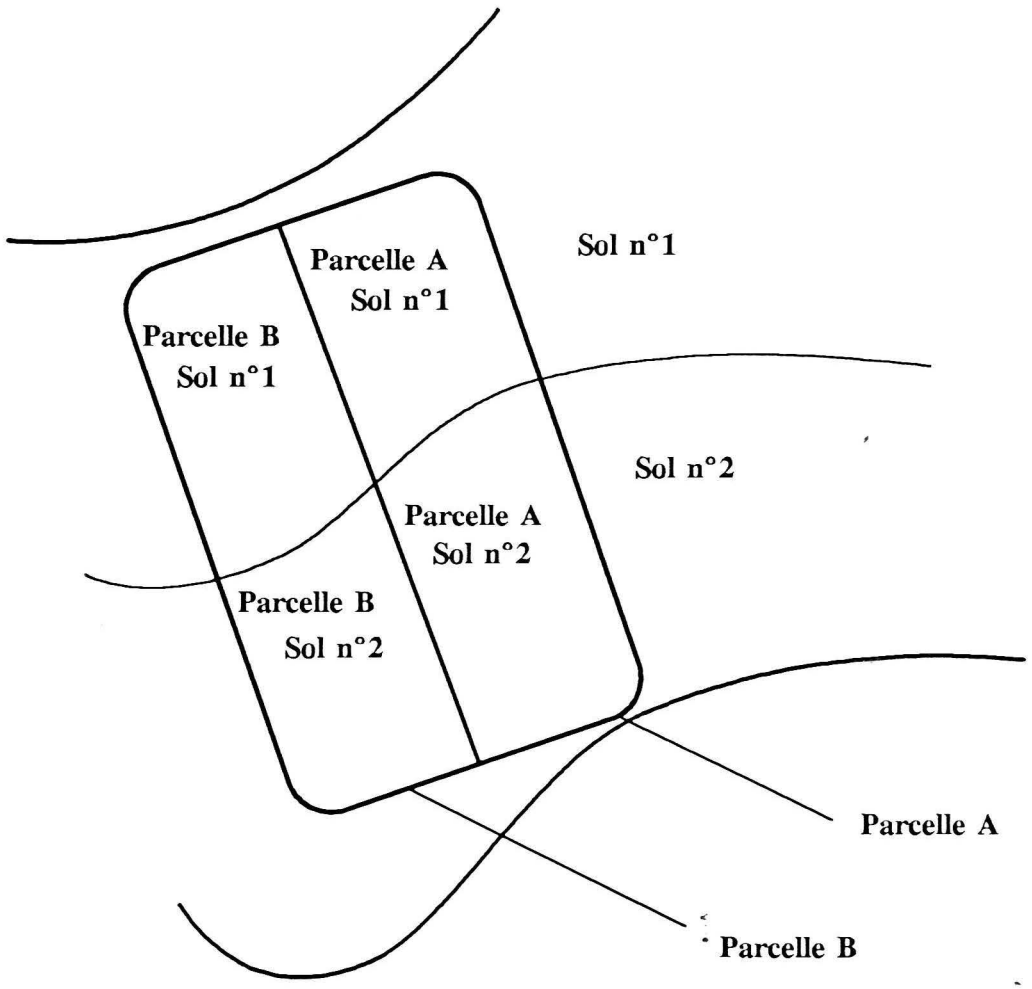


Figure n°8: division de l'espace

"région").
Ceci pose donc de gros problèmes de gestion des données géographiques mais aussi attributaires. Afin de ne pas compliquer l'étude de façon excessive, nous ne prendrons pas en compte les évolutions des données géographiques au cours du temps.

2. Les données attributaires.

Les variables attributaires, qui sont attachées aux objets géographiques, peuvent aussi évoluer au cours du temps. Ces évolutions se présentent sous trois formes :

* sous la forme de tableaux : évolutions de variables simples au cours du temps,
* sous la forme de cubes de données : évolutions de tableaux de données au cours du temps,
* sous la forme de courbes. Les courbes peuvent se ramener sous la forme de tableaux par une opération de discrétisation.

L'évolution des variables attributaires au cours du temps sera envisagée lorsque les méthodes statistiques le permettent.

E. Les objectifs de l'étude.

Nous avons défini, dans l'introduction, le S.I.G. comme un outil d'aide à la décision. En effet, il peut être utilisé comme un moyen de gestion et de remise à jour rapide de cartes, cependant, ses capacités le désignent plus comme un outil d'analyse et de synthèse. Il permet, par exemple, de superposer des cartes thématiques en respectant les échelles, donc de faire une synthèse visuelle.

De nombreux problèmes sont susceptibles d'être étudiés à l'aide d'un S.I.G. Cet outil cherche à approcher la complexité du monde réel.

Actuellement, les S.I.G. reçoivent une quantité énorme de données géographiques et attributaires sans beaucoup de cohésion. Certaines données sont même parfois recueillies sans que l'objectif de l'étude qui les utilisera n'ait été défini. Une importante banque de données sera donc disponible, mais la démarche est inversée : il est nécessaire de définir le phénomène à étudier avant de choisir les variables qui seront susceptibles d'être utilisées.

Dans toute cette étude, nous nous placerons dans un objectif exploratoire. Nous ne chercherons pas à adapter ou utiliser un modèle particulier sur nos données. Ce choix, certes réducteur (il existe de nombreuses autres méthodes) ne nécessite pas l'utilisation d'hypothèses trop lourdes, et constitue une première approche logique des données.

Dans le cadre d'un S.I.G., un certain nombre de questions exploratoires peuvent se poser, une partie d'entre elles trouveront une réponse totale ou partielle par l'analyse des données. Les questions sont du type :

* quelle est l'influence d'un axe de communication, d'une ville, d'un port sur le développement d'une région,
* quelle est la raison du développement d'une région,

* en quoi deux régions sont-elles comparables,

* la géographie explique-t-elle partiellement ou en totalité les valeurs des variables attributaires,

* peut-on définir des zones homogènes en terme d'occupation des sols, de densité du réseau routier, etc,

* y-a-t-il une structure répétitive dans l'organisation de l'espace (exemple : parcellaires autour d'une ville : plus les parcelles sont éloignées de la ville, et plus leur taille augmente),

* etc.

Toutes ces questions peuvent se traduire du point de vue mathématique par les problématiques suivantes :

* comparaison entre objets physiques de même nature,

* comparaison de relations de même type entre objets physiques différents,

* comparaison de lieux,

* mise en évidence de corrélations,

* etc.

Pour mettre en évidence les possibilités offertes par l'analyse des données, il est nécessaire d'étudier, dans le détail, le fonctionnement et les objectifs des différentes méthodes statistiques qui sont susceptibles de répondre à cette problématique.

II. Structuration des données.

A. Préparation des données.

1. Choix des individus statistiques.

Le choix des individus aux sens statistique du terme doit être dicté par l'objectif de l'étude. Un individu est une entité :

* qui a un support physique : une ville, une route, un département, etc...
* sur laquelle un certain nombre d'informations (données attributaires) sont connues.

A un choix d'individus statistiques correspond un niveau d'étude : on peut choisir de travailler sur des villes pour faire des comparaisons entre elles, alors que l'on connaît toutes les informations pour travailler sur les arrondissements.

Lorsque le choix des individus statistiques a été convenablement réalisé en fonction des objectifs de l'étude, il est nécessaire de structurer les données pour l'analyse des données.

Le nombre total d'individus statistiques sera noté dans la suite de cette étude "n".

2. les données géographiques.

Nous avons extrait des données géographiques initiales (sous la forme de cartes), dans la première partie de cette étude, un certain nombre d'informations : des distances entre individus, et des relations topologiques. Il faut standardiser ces informations pour pouvoir les utiliser.

a. Les distances géographiques.

Les distances géographiques observées sur les cartes, ou déduites des variables géographiques, devront être des pré-dissimilarités au sens mathématique du terme ($d(a,b) \in \mathbb{R}$, $d(a,b) = d(b,a)$, $d(a,a) = 0$) pour pouvoir être utilisées en analyse des données. Cette limite est une restriction technique. En fait, la réalité des distances géographiques impose $d(a,b) \in \mathbb{R}^+$ ($d(a,b) \geq 0$), donc nous travaillerons sur des dissimilarités mathématiques.

Ces distances géographiques nous définissent donc un premier type de dissimilarité D_{obs} de dimension n*n (nombre total d'individus statistiques) du type :

$$D_{obs} = \begin{bmatrix} & d_{ij} & \\ & & \end{bmatrix}$$

D_{obs} est une matrice symétrique positive à diagonale nulle. d_{ij} est la dissimilarité entre les individus i et j. Par exemple, Si les individus statistiques sont des villes, d_{ij} sera la distance géographique entre la ville i et la ville j (se reporter au paragraphe I.A.4.a.).

b. Les relations topologiques.

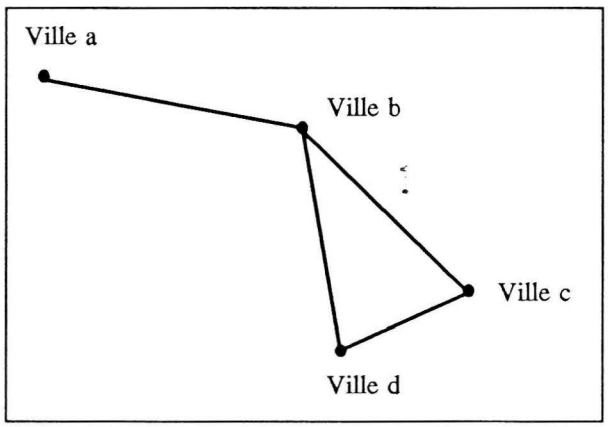
Pour la comparaison d'objets de même nature, la seule relation topologique que nous allons utiliser est la relation de voisinage. A partir de cette notion, il faut définir une matrice qui soit utilisable en analyse des données : une matrice U de voisinage de dimension n*n telle que :

* $U_{ij} = 0$ si (i = j) ou si (i et j) ne sont pas voisins,
* $U_{ij} = 1$ si ($i \neq j$) et (i et j sont voisins).

* Voisinage sur un graphe :

Comme nous l'avons vu dans la première partie, les graphes (représentations des réseaux) définissent un voisinage entre noeuds de graphes.

Par exemple pour 4 villes a,b,c,d ($n = 4$, $i \in \{a,b,c,d\}$), on dira que deux villes sont voisines si elles sont reliées par une route nationale. Pour le graphe suivant :

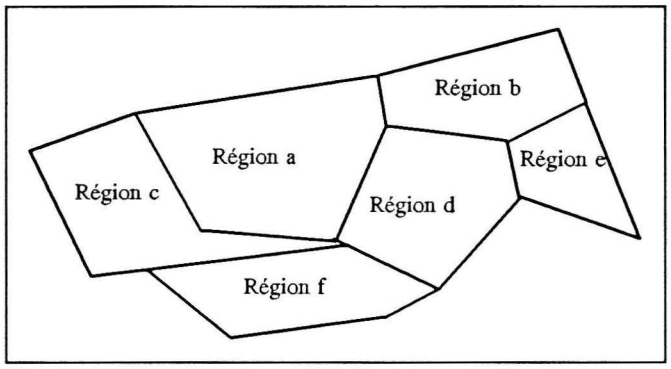


la matrice de voisinage U est :

$$U = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

* Voisinages par frontière commune :

La relation de voisinage au sens "une frontière commune", donne de la même manière une matrice U de voisinage. Par exemple, si l'on considère les six régions du graphe suivant,



on obtient la matrice de voisinage :

$$U = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \end{bmatrix}$$

On remarque bien, sur cet exemple, que la notion de voisinage est une façon imprécise de retranscrire la complexité d'une carte. La région a n'est pas voisine de la région f , de la même façon que la région c et la région e ne le sont pas. Un codage en 0 ou 1 semble donc un peu grossier pour être utilisé dans l'étude de données géographiques. Pourtant, ce codage autorise l'intégration de la notion de voisinage dans l'analyse des données, c'est pour cela que nous l'utiliserons.

La notion de voisinage est très souple, Il est possible de définir un voisinage à partir d'une distance géographique : deux individus seront voisins si la distance qui les sépare est inférieure à un nombre de référence fixé arbitrairement.

Il est possible, sans modifier la méthode développée au III.B., de définir un voisinage plus souple avec des chiffres compris entre 0 et 1. Cela revient à valuer le graphe de voisinage. Dans le cadre de cette étude, cela peut être fait en ramenant une distance géographique observée entre 0 et 1.

c. Variables géographiques simples.

Les données géographiques pourront aussi être utilisées sous la forme d'un tableau croisant les individus (objets géographiques) avec des variables simples. Le choix de ces variables est ouvert, et peut notament être les coordonnées des individus dans l'espace. Ce tableau de variables géographiques sera noté Y dans la suite de l'étude.

3. Les données attributaires.

Comme pour les données géographiques, il faut organiser les variables attributaires pour pouvoir les exploiter dans une analyse.

a. variables attributaires simples.

Le croisement des individus statistiques avec les variables attributaires simples qui auront été choisies pour l'étude donne un tableau de données qui est sous la forme d'une matrice :

$$X = \begin{bmatrix} & X_{ij} & \\ & & \end{bmatrix}$$

L'élément X_{ij} de cette matrice est la valeur de la variable j pour l'individu i. n sera le nombre total d'individus, et p le nombre total de variables :

$$\begin{matrix} i \in \{1,\ldots,n\} \\ j \in \{1,\ldots,p\} \end{matrix}$$

X_i est l'individu i défini par la ligne i du tableau X.

$$X_i = \begin{bmatrix} X_{i1} \\ . \\ . \\ X_{ij} \\ . \\ . \\ X_{ip} \end{bmatrix}$$

X_i a p coordonnées, il appartient à un espace vectoriel F de dimension inférieure ou égale à p, appelé "espace des individus". Il est possible de définir une notion de distance, plus ou moins complexe, entre individus.
La distance "classique" peut être utilisée : $d^2(X_i,X_l) = \Sigma_j (X_{ij} - X_{lj})^2$, soit avec la notation vectorielle $d^2(X_i,X_l) = (X_i - X_l)^t$ Id $(X_i - X_l)$ (Id est la matrice identité de dimension p*p).
Il est aussi possible de définir des distances plus complexes à partir, d'une matrice Q symétrique définie positive, non diagonale (de dimension p*p) :

$$d^2(X_i,X_l) = (X_i - X_l)^t \, Q \, (X_i - X_l)$$

(F,Q) est alors un espace euclidien (les individus ont une représentation euclidienne dans F), on dira alors que la distance ainsi définie est euclidienne.

b. Tableaux croisés.

A chacun des individus statistiques correspond un tableau croisé de données. On a donc n tableaux exactement du même type. On note k l'indice des individus statistiques : k $\in$ { 1 , ... , n }, i l'indice des lignes : i $\in$ { 1, ... , q } (q nombre total de lignes), et j l'indice des colonnes (variables simples) j $\in$ { 1 , ... , p }.

$$X^1 = \begin{bmatrix} \\ X_{ij}^1 \\ \\ \end{bmatrix} , \; X^2 = \begin{bmatrix} \\ X_{ij}^2 \\ \\ \end{bmatrix} , \; \ldots , \; X^n = \begin{bmatrix} \\ X_{ij}^n \\ \\ \end{bmatrix}$$

c. Courbes.

A chacun des individus statistiques correspond une ou plusieurs courbes. Il faut discrétiser ces courbes : on choisit arbitrairement des instants sur la courbe et l'on relève les valeurs de la ou des variables considérées. Entre les individus statistiques, la méthode de discrétisation ne varie pas, et les instants restent identiques.
Si on a plusieurs courbes on obtient un tableau et on se retrouve dans le cas précédent. Si on a une seule courbe, on se retrouve dans le premier cas : à chaque individu correspond un vecteur de variables : ce sont des "variables*temps". On peut garder la notion d'ordre (chronologique) entre variables en munissant le tout d'une matrice de voisinage (consulter A. Carlier [3 et 4] : Analyse Factorielle des Evolutions A.F.E.).

B. Quelles méthodes pour quels types de données.

A chaque type de données peut s'adapter, dans le meilleur des cas, une méthode exploratoire particulière. Les types de données attributaires et géographiques se combinent et forment des contraintes techniques qui limitent le choix des analyses utilisables. *Dans cette étude nous ne nous intéresseront pas aux analyses qui ne travaillent que sur un unique tableau* : les données attributaires et géographiques sont absolument indissociables.

1. Adéquation des méthodes aux types des données.

Le tableau ci-dessous résume l'adéquation types de données - méthodes.

Données géographiques	**Données attributaires**	**Méthodes**	**P.**
Dissimilarité	Variables simples ou croisées	Méthodes en cours de développement	/
	Distance euclidienne calculée	Comparaison de dissimilarités	37
Variables simples (distance euclidienne)	Variables simples	ACPVI / Analyse Canonique	25
	Tableaux croisés	STATIS ou AFM modifiée	/
Graphe de voisinage	Variables simples	Analyse Locale ou Globale	19
	Tableaux croisés	STATIS ou AFM modifiée	30

La case grisée au centre du tableau signifie qu'il y a adéquation des méthodes aux données, mais pas obligatoirement des objectifs des méthodes aux objectifs des études envisagées.

On retrouve toutes les situations abordées dans la première partie de cette étude. Les données géographiques sont :

* des dissimilarités : distances géographiques entre individus,
* des variables simples qui définissent une distance euclidienne entre les individus.

Dans notre problématique, la seule distance euclidienne entre individus, qui existe, est celle dite "à vol d'oiseau". La méthode la plus simple, pour l'obtenir, est de partir de variables simples : X, Y, Z et de la calculer. Or, dans les méthodes statistiques, il est plus facile de manipuler des variables, donc nous avons choisi de travailler sur ces variables simples.

* des graphes de voisinages.

Les données attributaires sont des variables simples ou croisées, et éventuellement une distance euclidienne qui sera définie à partir de ces variables.

Dans le cadre euclidien, l'adéquation données - méthodes se fait bien, c'est en effet le domaine de prédilection de l'analyse des données, mais hors du cadre euclidien les méthodes sont inexistantes ou en cours de développement.

2. Adéquation des objectifs.

Chacune de ces méthodes a un objectif précis, et a été développée dans un cadre qui ne correspond pas obligatoirement à celui de notre problématique.

a. Comparaison de dissimilarités.

Cette méthode permet de savoir si les deux dissimilarités (distance géographique et distance issue des variables attributaires) sont proches ou éloignées. On ne peut pas par cette méthode déterminer les fondements de cette situation.

Il est possible de définir et de calculer un indice de distance entre dissimilarités, ce qui permettrait de comparer et de classer différentes situations observées.

b. ACPVI et AC.

L'ACPVI et L'Analyse Canonique sont deux méthodes distinctes avec des objectifs différents.

Dans l'ACPVI, les rôles des deux tableaux étudiés (X variables attributaires, et Y variables géographiques) sont nettement *dissymétriques* : les variables géographiques expliquent les variables attributaires. L'objectif de cette méthode est de définir comment et dans quelle mesure les variables géographiques expliquent les variables attributaires. D'un

point de vue plus mathématique, on recherche les composantes principales, combinaisons linéaires de X et de Y qui maximisent l'inertie de X.

L'analyse canonique fait jouer aux deux tableaux le même rôle. On ne sait pas *a priori* qui explique quoi. L'objectif de cette méthode est de déterminer ce qu'ont en commun les deux tableaux. On recherche les combinaisons linéaires des variables de X et de Y qui sont au maximum corrélées.

Le choix des variables explicatives est très important dans la mise en oeuvre de ces méthodes. Si on utilise les variables qui définissent les positions des objets géographiques dans l'espace, cela signifie que l'on recherche un gradient linéaire, ce qui limite l'efficacité des méthodes. Ce problème est repris dans le paragraphe IV. sur un exemple.

c. Analyses Locales et Globales.

L'objectif de l'analyse locale (globale) est de trouver les composantes principales qui varient le plus (le moins) entre voisins. Les deux analyses conduisent à des résultats qui se complètent.

L'analyse globale permet de dégager de larges tendances en éliminant l'influence des variations locales. Les composantes principales de l'analyse globales sont les variables (combinaisons linéaires des variables de départ) qui présentent les meilleures caractéristiques pour être cartographiées. Cette méthode permet de mesurer la façon dont les variables de départ sont susceptibles d'être cartographiées grâce aux corrélations avec ces composantes principales. On étudie, donc, la manière avec laquelle les données attributaires peuvent être ou ne pas être expliquées par les données géographiques. On obtient les combinaisons linéaires qui sont, d'une façon optimale, retranscriptibles sur la carte de départ. On a, de la même manière, les combinaisons linéaires des variables de départ qui sont le moins cartographiable.

l'analyse locale étudie les variations entre voisins, elle permet de mettre en évidence des anomalies locales.

d. Statis (ou AFM) modifiée.

Les objectifs de ces méthodes sont les mêmes que pour les analyses Globales et Locales puisqu'elles sont en partie basées sur leurs principes. C'est une adaptation des analyses Locales ou Globales à l'analyse conjointe de plusieurs tableaux. Les différences entre STATIS et l'AFM sont envisagées dans la troisième partie.

Afin de cerner les capacités des différentes méthodes, nous allons étudier, dans une troisième partie, leurs fonctionnements mathématiques. Cette troisième partie pourra paraître un peu théorique, mais est absolument nécessaire pour une meilleure compréhension des mécanismes et de leurs utilités.

III. Mécanismes mathématiques des méthodes statistiques.

A. Variables simples et graphe de voisinage.

Ce type d'analyse permet d'exploiter un tableau individus-variables quantitatives X associé à une matrice de voisinage U. La matrice X définit un espace euclidien de dimension élevée dans lequel sont placés les individus. Le principe est d'obtenir une représentation du nuage des individus qui déforme le moins possible la réalité. Le graphe de voisinage permet d'introduire une décomposition de l'inertie totale qui autorise la prise en compte de critères originaux de "moindre déformation".

1. Les notations utilisées:

Soit (X,Q,D_n) le triplet statistique classique dont l'étude donnerait une ACP généralisée.

X est de dimension n*p: n individus et p variables.
On notera X_i l'individu i, $X_i \in E_p$ de base $\{e_j\}$ et de dimension p. E_p^* est l'espace dual de E_p, de base $\{e_j^*\}$. X^j est la variable j, $X^j \in F_n$ de base $\{f_i\}$ et de dimension n. F_n^* est l'espace dual de F_n, de base $\{f_i^*\}$.
Q est la métrique de l'espace des individus (E_p): $d^2(X_i,X_j) = (X_i-X_j)^t\, Q\, (X_i-X_j)$
D_n est la métrique de l'espace des variables (F_n): elle est toujours diagonale et contient le poids de chaque individu: $D_n = \mathrm{diag}[P_i]$.
U est la matrice de voisinage d'élément $m_{i,l}$. $m_{i,l} = 1$ si les individus i et l sont voisins, $m_{i,l} = 0$ sinon. On pose $\Sigma_{(i,j)}\, m_{ij} = 2m$, donc m est le nombre de couples de voisins.
On notera P_i^* la somme des poids des voisins de l'individu i: $P_i^* = \sum_l m_{i,l}\, P_l$ et D_n^* la matrice diagonale des P_i^*: $D_n^* = \mathrm{diag}[P_i^*]$.
$X_{v(i)}$ sera la moyenne des voisins de i: $X_{v(i)} = (\sum_l m_{i,l}\, P_l\, X_l)/ P_i^*$.

2. Mise en évidence d'un phénomène spatial.

Pour déterminer si les variables qui ont été observées dans différents lieux de l'espace géographique, ont une structure spatiale, deux tests peuvent être utilisés. Ils sont exposés dans le livre Cliff A.D. et Ord J.K. [7]. Ces deux tests reposent sur des indices : l'indice de Geary et l'indice de Moran. L'indice de Geary compare la variance locale à la variance de l'ensemble des couples, et l'indice de Moran indique l'importance de l'autocorrélation positive (faiblesse de la variance locale).

Ces indices sont successivement maximisés par les analyses que nous allons étudier: l'analyse locale maximise l'indice de Geary, et l'analyse globale maximise l'indice de Moran.

3. Décomposition de l'inertie.

Par définition, pour une matrice X centrée (l'origine de E_p correspond au centre de gravité du nuage de points), l'inertie totale (I_t) du nuage de points est:

$$I_t = \sum_{i=1}^{n} P_i \| X_i \|_Q^2$$

On peut aussi écrire l'inertie totale sous la forme:

$$2 * I_t = \sum_{i=1}^{n} \sum_{l=1}^{n} P_i P_l \| X_i - X_l \|_Q^2$$

L'inertie totale se décompose en inertie entre voisins (l'inertie locale) et en inertie entre non voisins (inertie globale):

$$2 * I_t = \sum_{i=1}^{n} \sum_{l=1}^{n} m_{i,l} P_i P_l \| X_i - X_l \|_Q^2 + \sum_{i=1}^{n} \sum_{l=1}^{n} (1 - m_{i,l}) P_i P_l \| X_i - X_l \|_Q^2$$

Dans un premier temps, c'est l'inertie locale qui va nous intéresser: on va chercher à maximiser cette inertie afin de trouver les combinaisons linéaires des variables initiales qui varient le plus entre voisins.

$$2 * I_L = \sum_{i=1}^{n} \sum_{l=1}^{n} m_{i,l} P_i P_l \| X_i - X_l \|_Q^2$$

$$2 * I_L = \sum_{i=1}^{n} \sum_{l=1}^{n} m_{i,l} P_i P_l \left[\| X_i \|_Q^2 + \| X_l \|_Q^2 - 2 * (X_i \mid X_l)_Q \right]$$

$$2 * I_L = 2 * \left(\sum_{i=1}^{n} \left(\sum_{l=1}^{n} m_{i,l} P_l \right) P_i \| X_i \|_Q^2 - \sum_{i=1}^{n} \sum_{l=1}^{n} m_{i,l} P_i P_l (X_i \mid X_l)_Q \right)$$

$$I_L = \sum_{i=1}^{n} P_i P_i^* \| X_i \|_Q^2 - \sum_{i=1}^{n} p_i \left(X_i \mid \sum_{l=1}^{n} m_{i,l} P_l X_l \right)_Q$$

$$I_L = \sum_{i=1}^{n} P_i P_i^* \| X_i \|_Q^2 - \sum_{i=1}^{n} P_i P_i^* \left(X_i \mid \frac{\sum_{l=1}^{n} m_{i,l} P_l X_l}{P_i^*} \right)_Q$$

$$I_L = \sum_{i=1}^{n} P_i P_i^* \| X_i \|_Q^2 - \sum_{i=1}^{n} P_i P_i^* \left(X_i \mid X_{v(i)} \right)_Q$$

$$\underline{I_L = \sum_{i=1}^{n} P_i P_i^* \left(X_i \mid X_i - X_{v(i)} \right)_Q}$$

On peut encore écrire l'inertie locale sous la forme de la trace d'un opérateur:

$$I_l = \sum_{i=1}^{n} P_i P_i^* \langle Q(X_i) \ , \ X_i - X_{v(i)} \rangle$$

$$I_l = \sum_{i=1}^{n} \langle Q(X_i) \ , \ P_i \left(P_i^* X_i - P_i^* X_{v(i)} \right) \rangle$$

$$I_l = \sum_{i=1}^{n} \langle Q X^t (f_i^*) \ , \ P_i P_i^* X^t (f_i^*) - \sum_{l=1}^{n} m_{i,l} P_l P_i X^t (f_l^*) \rangle$$

$$I_l = \sum_{i=1}^{n} \langle Q X^t (f_i^*) \ , \ P_i^* X^t D_n (f_i) - \sum_{l=1}^{n} m_{i,l} P_l X^t D_n (f_l) \rangle$$

$$I_l = \sum_{i=1}^{n} \langle Q X^t (f_i^*) \ , \ X^t D_n D_n^* (f_i) - X^t D_n U D_n (f_i) \rangle$$

$$I_l = \sum_{i=1}^{n} \langle Q X^t (f_i^*) \ , \ X^t D_n (D_n^* - U D_n) (f_i) \rangle$$

$$I_l = \sum_{i=1}^{n} \langle (f_i^*) \ , \ X Q X^t D_n (D_n^* - U D_n) (f_i) \rangle$$

$$I_l = trace \left(X Q X^t D_n (D_n^* - U D_n) \right)$$

Soit, après réarrangement:

$$\underline{I_l = trace\ (X^t D_n (D_n^* - UD_n) XQ)}$$

On reconnaît ainsi une expression proche de celle de l'inertie totale:

$$I_t = trace\ (X^t D_n XQ) = trace\ (VQ)$$

Seule, une matrice L modifie cette expression: $L = D_n^* - UD_n$.

Propriétés de L:

$(D_n^* - UD_n)$ définit un opérateur de F: $L \in \mathcal{L}(F)$
L est D_n symétrique:

$$(D_n^* - UD_n)^t D_n = (D_n^{*\,t} - D_n^t U^t) D_n = (D_n^* - D_n U) D_n$$

$$or \quad UD_n = D_n U$$

$$donc \quad (D_n^* - UD_n)^t D_n = (D_n^* - UD_n) D_n = D_n (D_n^* - UD_n)$$

(Les trois matrices U, D_n, D_n^* sont symétriques)

Inertie globale:

Nous avons étudié l'inertie locale, il est possible de faire de même pour l'inertie globale. La démarche est du même type, et on obtient l'expression suivante:

$$I_g = trace\ (X^t D_n\ (I_n - D_n - D_n^* - (1_{n*n} - U) D_n)\ XQ)$$

On retrouve ici aussi un opérateur D_n symétrique $G = I_n - D_n - D_n^* - (1 - U)D_n$

4. Réalisation d'une ACP locale.

En ACP classique, les axes principaux sont ceux qui maximisent l'inertie totale du nuage de points, mais pour une analyse locale on utilisera l'inertie locale: entre voisins.

Comme nous l'avons vu, l'inertie locale s'écrit sous la forme:

$$I_l = trace\ (X^t D_n (D_n^* - UD_n) XQ)$$

* Montrons que $u = X^t D_n (D_n^* - UD_n) X Q$ est un opérateur auto-adjoint de E_p:
u est une application linéaire de E_p dans E_p donc pour que u soit auto-adjoint il faut et il suffit que $Qu = u^tQ$.

$$\begin{aligned} u^tQ &= (X^t D_n (D_n^* - UD_n) XQ)^t Q \\ &= (QX^t (D_n^* - UD_n)^t D_n X) Q \\ &= Q(X^t D_n (D_n^* - UD_n) XQ) \\ &= Qu \end{aligned}$$

* On est donc dans le domaine de validité du théorème de Pearson: l'inertie locale projetée dans un espace de dimension k sera maximum dans l'espace engendré par les k premier axes obtenus par la décomposition spectrale de u.

Réaliser une ACP locale revient donc à diagonaliser le triplet statistique (X,Q,D_nL).

5. Réalisation d'une ACP globale.

L'objectif de cette démarche est de trouver des combinaisons linéaires des variables de départ qui varient peu entre voisins. Ces variables synthétiques ont la propriété de pouvoir être cartographiées.

La première idée est de travailler sur l'inertie globale (entre non voisins): réaliser l'ACP du triplet $(X,Q,(D_nG))$. La limite de cette démarche est que très souvent il y a beaucoup plus de non voisins que de voisins. En conséquence, réaliser l'ACP du triplet précédant revient, à peu de variations près, à faire une ACP classique. S'il y a un équilibre entre voisins et non voisins, cette démarche est alors tout à fait valable.

Une autre solution est de rechercher des combinaisons linéaires des variables de départ qui maximisent un indice d'auto-corrélation spatiale. L'indice utilisé est cèlui de Moran. Sous une forme matricielle, le numérateur de l'indice de Moran s'écrit:

$$X^tPX \quad avec \quad P = \frac{1}{2m} U$$

Si l'on tente de diagonaliser directement ce numérateur, comme le proposait D. Wartenberg [19], on se heurte à un problème de signe des racines. Une solution est apportée par D. Chessel et R. Sabatier [6]. Elle nécessite qu'il n'y ait pas de poids établis sur les individus statistiques. C'est en effet sur les poids que vient la modification: les poids uniformes (1/n) sont remplacés par:

$$P_i = \frac{\sum_{l=1}^{n} m_{(i,l)}}{2m} \quad avec \quad 2m = \sum_{i=1}^{n}\sum_{l=1}^{n} m_{(i,l)}$$

Cela signifie que plus un point a de voisins, plus son poids (donc sa contribution à l'inertie) est important.

Grâce à ces modifications (modification de D et centrage de X pour cette nouvelle pondération), l'expression des différentes inerties (locale et globale) change:

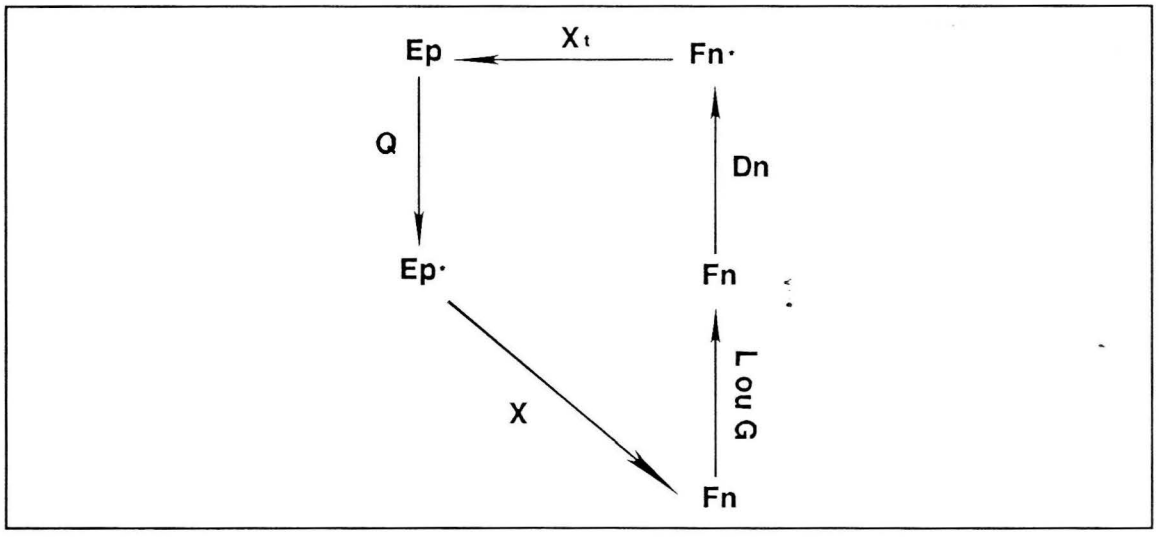
$$I_l = trace\ (X^t(D - P)XQ) \quad et \quad I_G = trace\ (X^tPXQ)$$

Il y a alors concordance entre le numérateur de l'indice de Moran et un élément de la décomposition de la variance de X:

$$X^tDX = X^t(D - P)X + X^tPX$$

La décomposition spectrale de la matrice X^t P X Q donne alors les composantes principales qui seront les plus facilement cartographiables.

L'introduction d'opérateurs D_n symétriques (L ou G) modifie le schéma classique de dualité de la façon suivante:



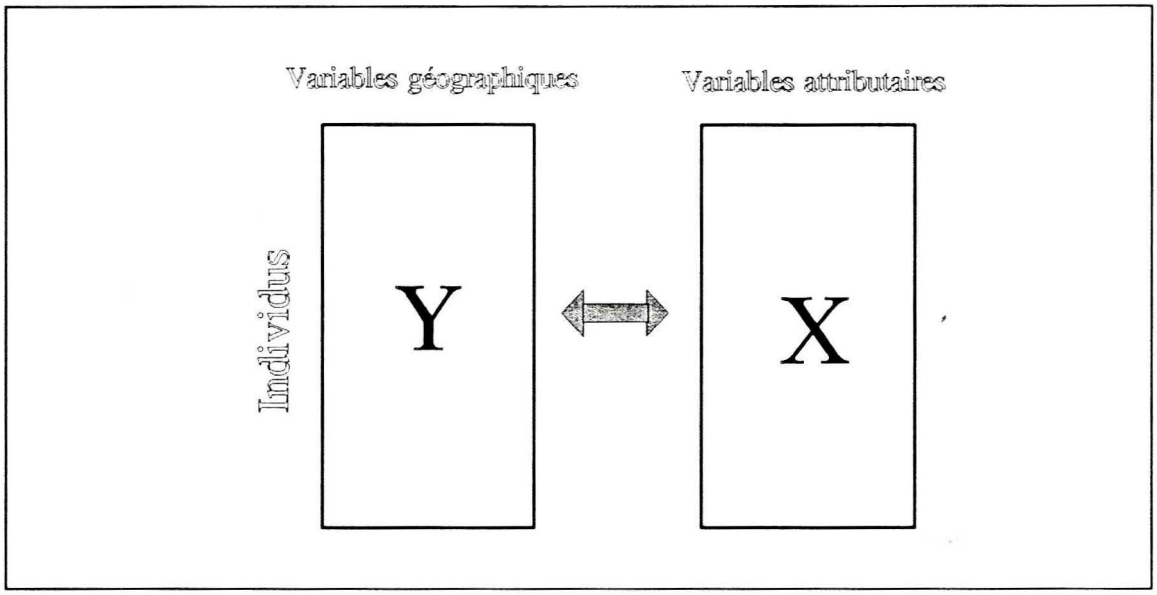
L'opérateur auto-adjoint de E_p à décomposer est:

* X^t D_n L X Q pour l'analyse locale,
* X^t D_n G X Q pour l'analyse globale.

B. Variables attributaires et géographiques simples.

Les deux types de données sont sous la forme d'un tableau objets géographiques * variables (attributaires ou géographiques). Les données géographiques sont sous cette forme lorsque l'on désire prendre en compte la distance "à vol d'oiseau" entre les supports physiques des individus statistiques. Cette situation se rencontre lors de l'étude de phénomènes naturels qui contournent les contraintes matérielles imposées par les infrastructures humaines. Les données géographiques seront sous la forme de coordonnées telles que longitude, latitude, altitude, ou encore les coordonnées de Lambert.

Nous appellerons X le tableau des données attributaire, et Y le tableau des données géographiques.



Deux stratégies sont possibles: soit les deux tableaux X et Y jouent un rôle dissymétrique soit ils jouent un rôle symétrique. Deux analyses distinctes répondent donc à ces deux problématiques: dans le premier cas l'ACPVI (ou l'AFCVI), et dans le deuxième l'Analyse Canonique ou l'analyse de co - inertie.

1. Analyses dissymétriques.

Nous avons donc deux tableaux (individus * variables) X et Y qui ne jouent pas le même rôle. Le premier sera le tableau de variables à expliquer (variables d'intérêt), et le deuxième de variables explicatives (variables instrumentales). Il est parfois délicat de faire une différence fondamentale entre les rôles des données attributaires et des données géométriques. Toutefois, nous supposerons que les variables géographiques sont des variables explicatives (image d'une structure) qui permettent de comprendre les variables attributaires.

a. Analyse en Composantes Principales avec Variables Instrumentales.

L'article de référence traitant de l'ACPVI est celui de Sabatier R., Lebreton J.D., Chessel D.: [17]. Cette analyse s'utilise lorsque X et Y sont des tableaux de variables quantitatives.

(1) Principe et définition d'une ACPVI.

X est le tableau des variables d'intérêt, et Y celui des variables instrumentales. D sera la métrique sur l'espace des individus (généralement D = 1/n Id, mais elle peut être différente) et Q sera la métrique sur l'espace des variables de X (généralement Q = Id). X et Y devront être deux tableaux, de variables quantitatives, centrés et éventuellement réduits suivant la nature des données.

Le principe de l'ACPVI de (X,Q,D) par rapport à Y, est de rechercher les composantes principales, combinaisons linéaires des variables définies par X, mais aussi par Y, qui maximisent l'inertie du nuage des points définis par X. Ces composantes principales sont donc à la fois des combinaisons linéaires des variables définies par X et par Y, mais on ne mélange pas les variables: les deux expressions sont distinctes.

D'un point de vue plus mathématique, et par définition:

On appel ACPVI d'ordre k du triplet (X,Q,D) par rapport à Y , la recherche de la métrique Q_y associée au triplet (Y,Q_y,D) de tel sorte que:

$$\| W_x D - Y Q_y Y' D \|^2_{HS}$$

soit minimal.

Cette distance entre opérateurs WD n'est généralement pas nûlle, sauf si l'espace défini par X est un sous espace vectoriel de celui défini par Y.

(2) Réalisation d'une ACPVI.

La solution du problème précédant n'est pas unique, mais il est possible de démontrer que les résultats de l'ACPVI ne dépendent pas du choix de la métrique Q_y.
Une solution généralement utilisée est donnée par la métrique Q_y suivante:

$$on\ pose \begin{cases} S_{YX} = Y^t D X \\ S_{YY} = Y^t D Y \end{cases}$$

$$et\ Q_Y = S_{YY}^- S_{YX} Q S_{XY} S_{YY}^-$$

Ceci donne un opérateur WD à diagonaliser qui a pour expression:

$$\begin{aligned} YQ_YY^tD &= Y\,(Y^tDY)^-\,(Y^tDX)\,Q\,(X^tDY)\,(Y^tDY)^-\,Y^t\,D \\ &= [\,Y\,(Y^tDY)^-\,Y^tD]\;X\;Q\;X^t\;[\,DY\,(Y^tDY)^-\,Y^t\,]\;D \\ &= P_YXQ(P_YX)^tD \end{aligned}$$

Avec P_Y projecteur D-orthogonal sur l'espace défini par Y.

$$P_Y = Y(Y^tDY)^-Y^tD$$

On a donc équivalence, pour la représentation des individus (même plan principal), entre la diagonalisation du triplet (Y,Q_Y,D) et celle du triplet (P_YX,Q,D). Pour les variables, les représentations des deux triplets se complètent:

* pour le triplet (Y,Q_Y,D), on a les corrélations des variables de Y (variables instrumentales) avec les composantes principales,

* pour le triplet (P_YX,Q,D), on a les corrélations des variables de X avec les composantes principales.

Une aide à l'interprétation est l'inertie expliquée par le plan principal (ou les k premiers axes) de l'ACPVI. Cette inertie doit être comparée avec celle de l'ACP simple du triplet (X,Q,D) (elle est toujours inférieure à cette dernière). Ce rapport permet d'avoir une idée de l'efficacité de l'ACPVI. S'il est très faible, il est alors inutile de faire une interprétation du premier plan de l'ACPVI: les variables explicatives (variables instrumentales) n'ont que peu de rapports avec les variables à expliquer (variables d'intérêt).

Il est également possible de réaliser une ACPVI dite orthogonale, en diagonalisant le triplet $(P_Y^{\perp}X,Q,D)$, pour éventuellement mettre en évidence une structure non expliquée par les variables de Y.

b. Analyse Factorielle des correspondances avec variables instrumentales.

Cette analyse répond à la même attente que l'ACPVI. Les variables d'intérêt sont, cette fois ci, qualitatives, les variables instrumentales restent quantitatives. Le tableau, X, des variables d'intérêt correspond soit à des variables de présence - absence (en 0 et 1), soit à des variables qualitatives sous la forme d'un codage disjonctif complet. Le tableau, Y, des variables explicatives devra être centré et éventuellement réduit.

La première étape est de préparer le tableau de données X comme pour une AFC: Il faut que $\Sigma_i\Sigma_j X_{ij} = 1$ donc si ce n'est pas le cas, on divise tous les éléments de X par $\Sigma_i\Sigma_j X_{ij}$.

On calcule ensuite les sommes par lignes (vecteur D_i) et par colonnes (vecteur D_j) des éléments de X. Soit DI et DJ les matrices diagonales des vecteurs D_i et D_j.

$$DI_{n*n} = \begin{bmatrix} . & 0 \\ & . \\ & X_{i.} \\ & . \\ 0 & . \end{bmatrix} \qquad DJ_{p*p} = \begin{bmatrix} . & 0 \\ & . \\ & X_{.j} \\ & . \\ 0 & . \end{bmatrix}$$

X a n lignes (individus statistiques) et p colonnes (variables d'intérêt).
X est ensuite modifié en $X1 = DI^{-1} * X * DJ^{-1} - 1_{n*p}$
L'AFC de X correspond à la diagonalisation du triplet (X1,DJ,DI), et l'AFCVI à la diagonalisation du triplet (P_YX1,DJ,DI). L'interprétation de l'AFCVI reste la même que celle de l'ACPVI.

2. Analyses symétriques.

Les deux tableaux X et Y jouent le même rôle, il est impossible de dire *a priori* lequel des deux à structuré l'autre.

a. L'analyse canonique.

Une première méthode d'analyse simultanée de X et de Y est de travailler sur un unique tableau résultant de la fusion par ligne des tableaux initiaux. Il est possible d'adapter cette technique à des tableaux de natures différentes (un tableau quantitatif, et un tableau qualitatif) en travaillant sur les composantes principales issues des analyses simples préalablement réalisées (ACP et AFC).

L'analyse canonique est une méthode spécifiquement adaptée à l'étude simultanée de deux tableaux quantitatifs. Son objectif est de rechercher les combinaisons linéaires des tableaux de départ X et Y qui soient de corrélation maximale.

Cette problématique se traduit d'une façon simple avec les notations utilisées dans l'ACPVI: l'analyse canonique correspond à l'ACP du triplet statistique (T,S^-_{xx},S^-_{yy}), avec $T = Y^tD_nX$ (D_n est la métrique sur l'espace des individus). Cette diagonalisation nous donne les variables canoniques: les variables canoniques de l'espace défini par X sont les composantes principales et les variables canoniques de l'espace défini par Y sont les axes principaux.

La représentation des variables est double: les variables de X et de Y en fonction des variables canoniques de X, et les variables de X et de Y en fonctions de variables canoniques de Y. De la même manière, la représentation des individus (ce sont les mêmes pour X et Y) est double.

L'interprétation d'une analyse canonique est parfois difficile à réaliser. Dans notre problématique, cette analyse répond à une question importante: les données attributaires et géographiques traduisent-elles la même réalité ?, ou encore, les données attributaires sont-elles indépendantes des données géographiques ? (les données géographiques ont-elles structuré les données attributaires).

b. Analyse de co-inertie.

Chessel D., Mercier P. [5].

C'est une alternative, plus simple, à l'analyse canonique. Dans le cas de tableaux de variables quantitatives on retrouve l'analyse inter batterie proposée par Tucker en 1958. C'est la méthode préconisée par Chessel D. et Mercier P dans leur article "Couplage de triplets statistiques et liaisons espèces - environnement". Elle revient à faire l'analyse du triplet statistique (Y^tD_nX, D_p, D_q). D_n est la métrique sur les individus (commune au deux tableaux X et Y), D_p est la métrique sur les variables de X, et D_q est la métrique sur les variables de Y.

C. Tableaux croisés et graphes de voisinages.

Cornillon P.A. [8].

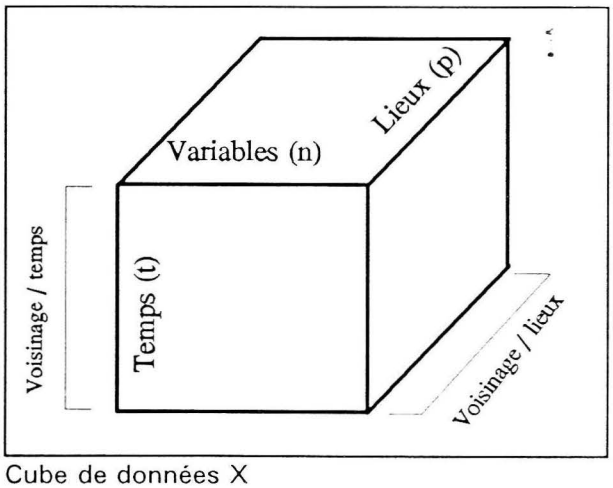
1. Les données.

Les individus statistiques (objets géographiques) sont définis par des données attributaires sous la forme de tableaux. Ces tableaux peuvent être (ou ne pas être) des variations de variables au cours du temps. Nous étudierons le cas le plus général (variation au cours du temps) en donnant en complément les résultats pour des tableaux de variables ne dépendant pas du temps.

Les données attributaires sont sous la forme d'un tableau à trois entrées: le temps, les variables et les objets géographiques (individus statistiques). Sur ces individus statistiques est défini un graphe de voisinage géographique. De plus, il est possible de définir un graphe de voisinage sur le temps. Ce dernier permet de prendre en compte l'ordre des données. Un voisinage sur le temps se définit généralement par : deux instants sont voisins si leur écart dans le temps est inférieur à un écart de référence.

Les notations seront, en conséquence, les suivantes:
X désigne le cube de données: t instants, p objets géographiques, n variables.
${}_kX^i_j$: est la valeur de la $k^{ième}$ variable à l'instant i, pour l'objet géographique j.
${}_kX$ est la matrice p*t correspondant à la variable k.
${}_kX^c$ est un vecteur colonne correspondant à la matrice ${}_kX$ dont les colonnes ont été mises les unes sous les autres.
χ est la matrice à n colonnes et p*t lignes correspondant à la mise des t tableaux, à p lignes et n colonnes, les uns sous les autres. $\chi_j = {}_jX^c$.
Q est la métrique sur l'espace des variables.
D_t est la métrique sur l'espace temps.
D_p est la métrique sur l'espace défini par les objets géographiques.
M_p est la matrice du graphe de voisinage sur les objets géographiques.
M_t est la matrice du graphe de voisinage sur le temps.

On peut, pour une meilleure compréhension du problème, matérialiser les données comme il suit:



Cube de données X

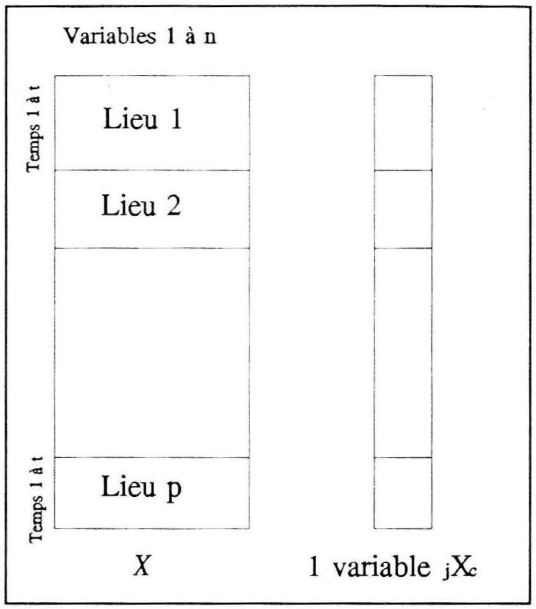


Tableau χ et variable ${}_iX^c$

On définit de plus le produit de Kronecker entre deux matrices A et B: ,

$$A \otimes B = \begin{bmatrix} a_{11} * B & a_{1n} * B \\ a_{p1} * B & a_{pn} * B \end{bmatrix}$$

2. Première méthode: ACP locale.

Cette méthode est directement dérivée de l'analyse locale que nous avons envisagée au paragraphe III.A. L'analyse se fait sur les variables vectorialisées ${}_kX^c$ de l'espace de dimension p*t.

L'opérateur de voisinage E devient:

* lorsque l'on a deux graphes de voisinages (espace géographique et temps)

$$E = D_p^* \otimes D_t^* - M_p D_p \otimes M_t D_t$$

avec D* la matrice diagonale des poids de voisinage (voir première partie),

* lorsque l'on a un seul graphe de voisinage (espace géographique):

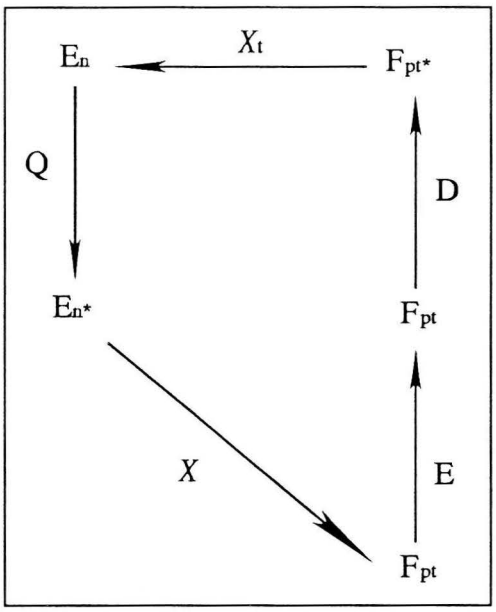
$$E = (n-1)/n\ D_p^* \otimes I_d - M_p D_p \otimes U D_t$$

avec I_d la matrice identité (à t lignes et t colonnes), et U la matrice ne contenant que des 0 sur la diagonale et des 1 partout ailleurs, U a t lignes et t colonnes.

Lorsqu'il n'y a pas de graphe de voisinage, on considère que chaque individu est voisin de tous les autres (sauf de lui même), ce qui explique la matrice U et le poids (n-1)/n issu de D_d^*

L'analyse locale du cube de données X revient alors à faire l'ACP du triplet (χ,Q,DE), avec D = $D_p \otimes D_t$ métrique de $\mathbb{R}^{pt}$ (le tableau χ doit être centré pour D). Cette ACP détermine les combinaisons linéaires des variables qui maximisent l'inertie locale du nuage des points (temps espace).

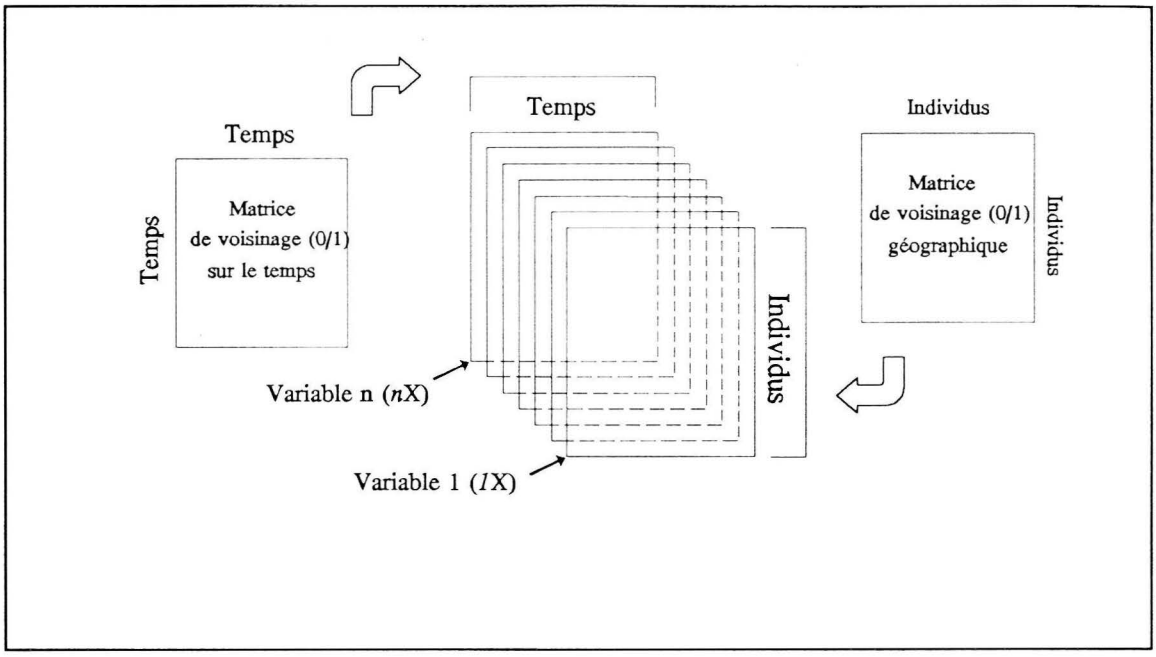
Le schéma de dualité devient le suivant:



Il est donc très proche de celui de l'analyse de tableaux simples avec graphe de voisinage, mais les espaces sont différents et χ est obtenu en juxtaposant les tableaux du cube de données X initial.

3. Deuxième méthode: STATIS modifié.

Dans son utilisation habituelle, un des intérêts de la méthode STATIS est de pouvoir travailler sur des tableaux dont une entrée peut être différente d'un tableau à l'autre. Dans notre problématique, nous avons introduit deux contraintes supplémentaires: les deux graphes de voisinage. Le voisinage sur les individus (objets géographiques) fait que les individus ne peuvent pas être différents d'un tableau à l'autre. De même, le voisinage sur le temps fait que l'entrée temps ne peut pas varier d'un tableau à l'autre. Ces voisinages imposent, de plus, que l'on découpe les données X en tableaux individus * temps. Donc l'entrée "variables" ne doit pas changer non plus. Il en résulte que nous sommes contraint de travailler sur un cube de données comme pour la première méthode.



Toutefois, si nous n'utilisons pas le voisinage sur le temps, il est possible que l'entrée temps ou variables soit différente d'un individus à l'autre. L'entrée identique sert alors à découper les données en tableaux sur lesquels ont calcule l'équivalent d'un W_k ($_kX$ D_t $_k{}^tX$).

Dans son rapport de DEA, P.A. Cornillon définit un semi produit scalaire "local" sur les éléments caractéristiques. Le choix des éléments caractéristiques de l'étude s'est porté sur les tableaux $_kX$. Les tableaux sont étudiés sans avoir besoin de les juxtaposer (χ) ni de calculer un W_k ou un V_k. Ce semi produit scalaire est défini par la semi métrique $H = D_pE_p \otimes D_tE_t$. On peut donc ainsi construire une matrice C de semi produits scalaires entre $_kX$ de dimension n*n et d'élément :

$$C_{l,k} = trace\ (\ {}_l^tX\ D_tE_t\ {}_kX\ D_pE_p\ ) = \langle\ {}_lX^c\ ,\ {}_kX^c\ \rangle_{D_pE_p \otimes D_tE_t}$$

Toutes les étapes de STATIS peuvent être réalisées pour une étude spatiale et temporelle des variations des $_kX$ en remplaçant le coefficient RV ou le produit scalaire entre étude par ce semi produit scalaire. Pour de plus amples détails sur la méthode STATIS, consulter le livre de Ch. Lavit [13].

a. Calcul de l'interstructure.

Dans cette étape, on diagonalise l'opérateur défini par CQ afin d'obtenir le plan principal dans lequel chacun des tableaux $_kX$ est représenté par un point. Ce plan permet de mettre en évidence une structure (variations locales et temporelles) entre les différentes variables qui ont été choisies.

Cette étape revient à faire l'ACP du triplet (χ,Q,H), ceci rapproche STATIS modifié de l'analyse locale définie précédemment (ACP du triplet (χ,Q,DE)). Les deux semi

métriques locales sont toutefois différentes:

* première méthode: $DE = D_p \otimes D_t * (D_p^* \otimes D_t^* - M_p D_p \otimes M_t D_t)$

* deuxième méthode: $H = D_p E_p \otimes D_t E_t$.

b. Calcul du compromis.

Le compromis se définit comme un résumé majoritaire des objets, c'est-à-dire, dans notre problématique, un tableau résumé ${}_mX$ des tableaux ${}_kX$. Ce compromis se définit, dans STATIS, à partir de la première valeur propre de l'opérateur CQ.

Par définition:

$$ {}_mX = \alpha \Sigma_k \Pi_k \lambda_{1k} {}_kX $$

Avec Π_k poids de chacune des études (généralement 1/n),
λ_{1k} $k^{ième}$ composante de la première valeur propre de CQ,
α coefficient de normalisation.

Le compromis revient donc à faire l'ACP du triplet $({}_mX, D_pE_p, D_tE_t)$. Le plan principal de cette ACP permet de situer les objets géographiques moyens (la moyenne est réalisée sur les variables) les uns par rapport aux autres. Les composantes principales sont des combinaisons linéaires des variables instants. La représentation des objets géographiques moyens sur un même plan factoriel permet de réaliser une comparaison entre eux du point de vue de leurs variations dans le temps.

c. Calcul de l'intrastructure (trajectoire).

L'intrastructure s'appuie sur le même plan principal que le compromis, elle correspond à la projection des différents ${}_kX$ en individus supplémentaires. Cette représentation permet d'expliquer les positions des points du compromis en fonction des variables, et donc, dans notre problématique, de trouver les variables au comportement local hors moyenne.

Le nom de trajectoire est, ici, impropre puisque l'indice k n'est pas basé sur le temps. Le choix de la façon de décomposer le cube de données X est en fait dicté par la présence des deux graphes de voisinages M_p et M_t.

Il faut noter que les tableaux ${}_kX$ sont projetés sur le plan principal de l'ACP du triplet $({}_mX, D_pE_p, D_tE_t)$. Leurs représentations n'ont pas obligatoirement de propriétés d'optimalité. Il faut donc, avant toute interprétation de ces projections, vérifier le pourcentage d'inertie représenté. L'intrastructure doit être prise comme une aide à l'interprétation et doit être manipulée avec prudence.

d. STATIS modifié avec un seul graphe de voisinage (géographique).

Les données se limitent donc à un cube de données X et un seul graphe de voisinage M_p lié aux objets géographiques. Nous continuerons à parler d' objets géographiques, de variables et de temps même si l'entrée du cube de données "temps" est différente d'une succession d'instants.

Il est toujours possible de décomposer le cube de données en tableaux correspondant chacun à une variable. Dans ce cas:

* l'interstructure est donné par l'ACP du triplet (χ,Q,H1), avec H1 = $D_pE_p \otimes D_t$,
* le compromis est donné par l'ACP du triplet ($_mX$,D_pE_p,D_t),
* l'intrastructure est donnée par la projection des tableaux $_kX$ sur le plan principal du compromis.

Il est maintenant possible de décomposer le cube de données en t tableaux X^i. L'interstructure est alors donnée par l'ACP du triplet ($\chi 2$,Q,H2), avec:

* $\chi 2$ tableau obtenu par la juxtaposition, en colonne, des p tableaux variables * temps,
* H2 = $D_pE_p \otimes Q$.

L'interstructure permet, ainsi, de représenter, dans un même plan factoriel, les t tableaux correspondant à l'entrée "temps".
Le compromis est l'ACP du triplet (X^m,D_pE_p,Q) et représente les objets géographiques moyens. L'intrastructure reste toujours la projection des tableaux X^i sur le plan principal du compromis.

4. Troisième méthode: modification de l'AFM.

L'Analyse Factorielle Multiple, définie par B. Escofier et J. Pagès [9], et STATIS sont deux méthodes qui suivent la même stratégie. Les variations entre ces deux méthodes ont toutefois leur importance.

* Au niveau de l'interstructure:

Le calcul de l'interstructure se fait de la même façon que pour la méthode STATIS, mais est seulement basé sur le produit scalaire entre WD. Le coefficient RV n'est pas utilisé. Les adaptations de STATIS pour prendre en compte les variations spatiales et temporelles sont donc tout à fait applicable à L'AFM (ACP du triplet (χ,Q,H)).

* Au niveau du compromis: pondération différente.

L'AFM nécessite la réalisation des ACP (ici locale) simples sur tous les tableaux $_kX$ issus de la décomposition du cube de données X. C'est l'inverse de la première valeur propre de chacune de ces ACP qui donne les poids dans le compromis des différents tableaux.
On définit donc $_mX$ différemment:

$$ {}_{m}X = \Sigma_{k} \, (1 / \lambda_{1k}) \, {}_{k}X $$

avec λ_{1k} première valeur propre de l'ACP du triplet $({}_{k}X, D_pE_p, D_tE_t)$

Le compromis de l'AFM sera donc l'ACP du triplet $({}_{m}X, D_pE_p, D_tE_t)$. Il faut noter que si les tableaux ${}_{k}X$ avaient eu un nombre de colonnes différent entre eux, le compromis ne pourrait pas se définir ainsi dans l'AFM. Cette situation ne pouvant pas se rencontrer dans le cadre de notre problématique, le calcul initial n'est pas ici détaillé (se reporter au livre cité précédemment).

La pondération de STATIS (composantes de la première valeur propre de la matrice CQ) a pour effet de réaliser "un compromis majoritaire": les études qui ne sont que faiblement corrélées avec l'axe principal pèsent moins dans le compromis et sont donc moins bien représentées que les études fortement corrélées.

La pondération de l'AFM réduit l'inertie maximum de chacune des études sur une direction à 1, donc l'inertie expliquée par les axes de l'ACP définissant le compromis n'est pas influencée par l'inertie initiale de chacune des études.

Les auteurs de la méthode définissent, de plus, des aides à l'interprétation basées sur les ACP simples des différentes études qu'il est nécessaire de réaliser pour construire une AFM. Il s'agit notamment des projections sur le plan principal du compromis, des objets et des variables des ACP séparées.

D. Distance observée et variables attributaires simples.

Dans cette situation, nous sortons du cadre euclidien car les données géographiques, sous la forme d'une distance quelconque, ne définissent pas un espace affine muni d'une métrique. L'identification d'une distance euclidienne peut se faire facilement par le calcul des coefficients de la matrice de Torgerson (voir Fichet B. [11]).

Le domaine non euclidien est le sujet de nombreuses recherches qui ont abouti à de nouvelles méthodes. Ces méthodes ne se diffusent que lentement, elles sont, en effet, soit encore trop nouvelles, soit trop "gourmandes" en calcul pour être appliquées actuellement. On peut, toutefois, citer deux méthodes:

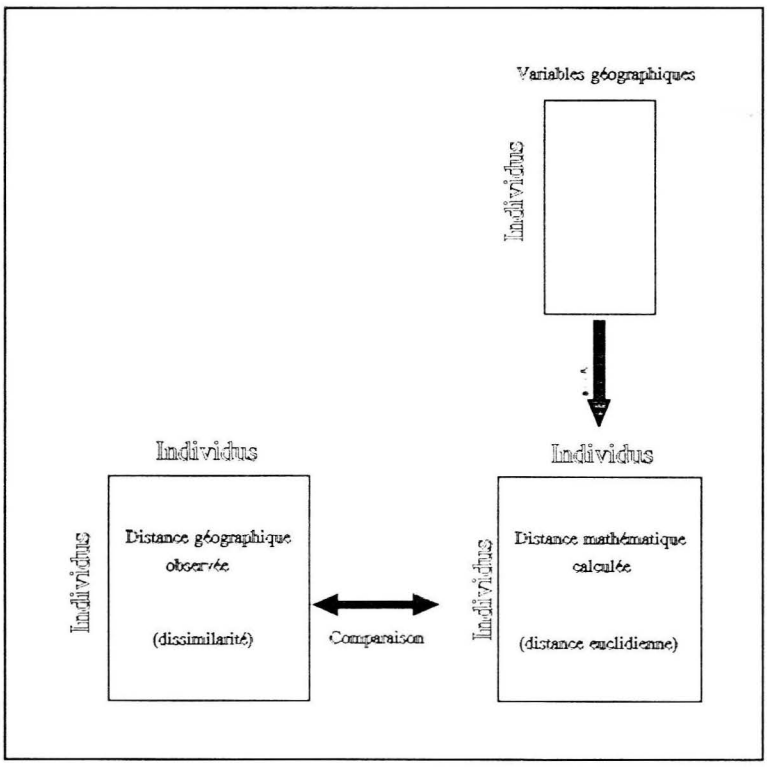
* l'A.C.P. en norme L_1 de Fichet B. (utile pour des distances de type Manhattan),
* la comparaison de dissimilarités.

C'est cette dernière méthode que nous avons choisi de développer dans cette partie.

1. Les notations utilisées.

Les données géographiques sont sous la forme d'une distance observée D_{obs} de dimension n*n. Du point de vue mathématique, D_{obs} n'est pas une distance, mais une dissimilarité.

Les données attributaires sont sous la forme d'un tableau de variables qui définissent un espace affine. Sur cet espace, il est possible de définir des distances (D_{cal}). C'est sous cette forme que l'information attributaire sera utilisée.



2. Démarche de la méthode.

A partir du tableau X des variables attributaires, on défini une distance euclidienne entre individus. Il est possible d'utiliser une métrique Q symétrique, non diagonale. Nous appélerons cette ditance D_{cal}.

Les résultats qui suivent, sont issus de l'article de B. Fichet [11].
Le cardinal de I (n) doit être strictement supérieur à 4.

Soit Δ l'espace vectoriel fini des fonctions définies par:

$$d : I^2 \text{ ----> } \mathbb{R}$$
$$\forall i \in I,\ d(i,i) = 0;$$
$$\forall (i,j) \in I^2,\ d(i,j) = d(j,i);$$
$$\forall (i,j) \in I^2,\ d(i,j) \geq 0.$$

Δ est l'espace vectoriel des dissimilarités, il est de dimension n(n-1)/2.
Les deux matrices D_{obs} et D_{cal} définissent deux fonctions éléments de Δ.

Une base de Δ est donnée par l'ensemble des bi - dissimilarités binaires de I.
Une dissimilarité binaire se définie à partir d'une partition de I en deux sous ensembles J et J_c ($J \cup J_c = I$, et $J \cap J_c = \varnothing$):

$$\forall (i,j) \in J,\ d(i,j) = 0,$$
$$\forall (i,j) \in J_c,\ d(i,j) = 0,$$
$$\forall i \in J \text{ et } \forall j \in J_c,\ d(i,j) = 1.$$

Une bi - dissimilarité binaire est une dissimilarité binaire dont un des deux ensembles ne possède que deux éléments. Si n est le cardinal de I, il y a n*(n-1)/2 bi - dissimilarités binaires qui forment une base de Δ. Nous noterons ces dissimilarités δ_{ij}, i et j sont les éléments de l'ensemble J (le cardinal de J est 2).

Soit d une dissimilarité quelconque de Δ, les coordonnées de d dans la base formée par les δ_{ij} sont donnés par l'expression suivante:

$$a^{ij} = 1/2\ [-d(i,j) + n/(n-4)*[d(i,.) + d(j,.)] - n^2/((n-2)(n-4))\ d(.,.)]$$

Dans cette expression, les notations utilisées sont $d(i,.) = \Sigma_v\ d(i,v)$ et $d(.,.) = \Sigma_i\ d(i,.)$.

Cette démarche permet d'avoir les coordonnées de toutes les dissimilarités de I dans un espace vectoriel commun Δ. Il est alors possible de calculer un indice de distance entre dissimilarités.

IV. Mise en oeuvre des méthodes.

Afin de tester l'efficacité des méthodes qui ont été exposées dans le paragraphe III, il est nécessaire de les faire fonctionner sur des exemples concrets. Le jeux de données initialement prévu pour cette étude, n'étant pas disponible, nous avons travaillé sur des exemples choisis au hasard, et non sélectionnés.

Parmi les méthodes présentées dans le paragraphe III, celles basées sur la notion de voisinage nous ont parues plus particulièrement intéressantes à étudier en raisons de leur relative nouveauté. L'ACPVI est une méthode déjà largement répandue, toutefois le cadre géographique étant particulier, nous avons tenté de montrer comment elle pouvait être exploitée. Les autres méthodes n'ont pas été illustrées en raison d'un manque de données adaptées.

Nous avons donc choisi de tenter, dans un premier temps, une comparaison de trois méthodes: l'ACP, l'ACP locale, et l'ACP globale au travers de leurs résultats sur des jeux de données et dans un deuxième temps, nous avons mis en oeuvre la méthode de l'ACPVI.

A. Comparaison ACP, ACP locale, ACP globale.

1. Premier exemple: les élections européennes à Paris.

Il s'agit des résultats des élections européennes du Dimanche 12 Juin 1994 qui ont été publiés par le journal Le Monde.

a. Présentation des données.

Dans cet exemple, les individus statistiques sont les 20 arrondissements de Paris. A partir de la carte de Paris (support géographique des individus statistiques), nous avons créé une matrice de voisinage à la manière de l'exemple théorique du paragraphe II.A.2.b.

Les variables étudiées représentent le nombre de votes accordés à chacune des listes présentes. Pour chacun des arrondissements, les votes sont donnés en pourcentage des suffrages exprimés. La liste "Europe pour tous" (EPT) a été éliminée du tableau, elle n'a, en effet, reçu aucun vote dans 14 arrondissements. Il faut noter que les scores obtenus par cette liste sont inférieurs aux erreurs commises lors du calcul des pourcentages. Dans aucun des arrondissements, la somme des pourcentages n'est égale exactement à 100%: elle est souvent inférieure, mais parfois supérieure.

Les données brutes sont dans le tableau n°1, la dernière colonne est la somme de toutes les précédentes.

Abréviations utilisées:

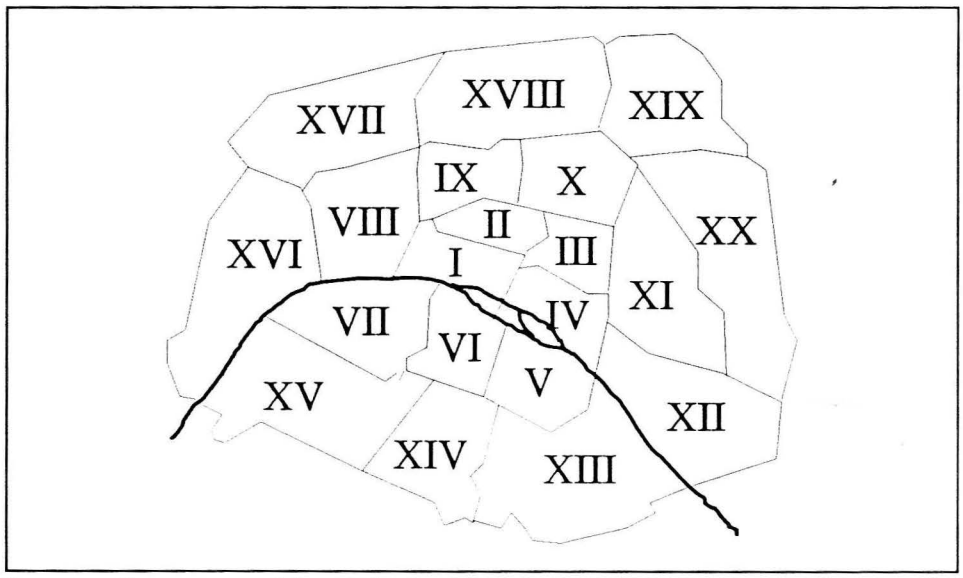
AutE: l'autre Europe (Philippe de Villiers),
AutP: l'autre politique (Jean Pierre Chevènement),

Données brutes des élections européennes

Arrond.	MAJ.	PS	Aut. E.	FN	MRG	PCF	Aut.P	Verts	GE	Saraj	LO	CPNT	Rég.	Out. Mer	PT	Deue	PLN	PVE	Emploi	EPT	somme
1	32.4	18.2	14.66	8.73	7.15	3.25	3.41	2.59	2.75	3.12	1.64	0.79	0.34	0	0.19	0.25	0.15	0.17	0.13	0	96.49
2	27.35	21.1	10.41	9.41	9.12	3.74	4.27	3.76	2.28	3.26	2.52	0.82	0.57	0.28	0.34	0.14	0.2	0.16	0.16	0	99.05
3	25.76	24.7	9.04	7.26	9.53	4.51	4.35	3.19	2.63	4.39	2.48	0.55	0.28	0.12	0.4	0.23	0.18	0.14	0.18	0	99.82
4	29.39	22.7	12.21	6.57	7.81	3.96	4.01	2.89	2.58	3.63	2	0.7	0.39	0.07	0.26	0.3	0.21	0.17	0.12	0	100.27
5	32.19	22.6	11.8	6.08	6.49	3.21	4.51	2.98	2.31	3.63	2.01	0.73	0.44	0.11	0.15	0.19	0.18	0.18	0.08	0	99.4
6	36.02	20.9	14.7	6.34	4.92	2.36	3.31	2.33	2.33	3.47	1.41	0.78	0.28	0.05	0.09	0.18	0.14	0.12	0.07	0.05	101.09
7	42.58	13.1	19.48	7.72	4.59	1.33	2.25	1.64	1.94	2.47	1.04	0.94	0.24	0.08	0.1	0.15	0.08	0.07	0.05	0	100.95
8	42.31	11.2	20.14	8.84	5.93	1.11	1.88	1.57	1.76	2.57	0.78	1.13	0.12	0.08	0.07	0.12	0.13	0.06	0.05	0.02	100.27
9	30.4	19.2	12.95	9.21	8.34	3.3	3.55	2.84	2.33	3.38	2.19	0.78	0.27	0.22	0.24	0.22	0.19	0.15	0.11	0	98.24
10	24.48	20.2	10.55	11.3	10.29	5.11	4.25	3.49	2.13	3.06	2.32	0.81	0.33	0.44	0.45	0.22	0.2	0.14	0.17	0	99.2
11	23.75	22.1	9.55	9.71	10.37	5.42	5.01	3.54	2.2	3.29	2.48	0.8	0.37	0.21	0.37	0.2	0.21	0.16	0.13	0	99.13
12	28.2	19	13.02	9.46	8.98	4.56	3.79	3.23	2.26	3	2.15	0.86	0.35	0.16	0.22	0.18	0.22	0.15	0.15	0	101.12
13	24.34	22.2	10.09	8.69	10.01	6.04	4.53	3.49	2.4	2.91	2.43	0.82	0.3	0.55	0.36	0.25	0.19	0.13	0.17	0	99.19
14	28.91	21.4	11.6	8.53	7.69	4.65	4.16	3.26	2.4	2.98	2.07	0.69	0.27	0.42	0.25	0.23	0.17	0.12	0.11	0	100.28
15	36.21	17.3	15.05	7.87	6.83	2.77	3.22	2.51	2.28	2.56	1.45	0.8	0.24	0	0.15	0.2	0.18	0.14	0.11	0	100.85
16	46.69	9.72	21.74	7.86	4.64	0.92	1.48	1.34	1.53	1.9	0.64	0.87	0.1	0	0.05	0.14	0.12	0.09	0.06	0.01	101.64
17	36.95	14	16.96	9.51	6.91	2.4	2.49	2.41	2.09	2.52	1.53	0.89	0.18	0.25	0.21	0.14	0.16	0.13	0.11	0.07	100.6
18	24.57	18.4	10.84	12.7	9.7	5.5	4.05	3.66	2.03	2.87	2.75	0.75	0.4	0.5	0.34	0.21	0.22	0.21	0.21	0	97.35
19	22.38	18.7	9.7	12	13.08	6.52	3.9	3.22	2.08	2.8	2.51	0.69	0.34	0.5	0.43	0.35	0.22	0.14	0.23	0.05	98.46
20	21.9	19.5	9.45	11.4	11.87	6.51	4.37	3.71	2.39	3.02	2.87	0.72	0.34	0.55	0.4	0.21	0.21	0.16	0.24	0.03	99.57

Tableau n° 1

CPNT: chasse pêche nature et traditions (André Goustat),
DEUE: démocrates pour les Etats Unis d'Europe (Armand Touati),
Empl: l'emploi d'abord (Gérard Touati),
EPT: Europe pour tous (Jean Aillaud),
FN: front national (Jean Marie Le Pen),
GE: génération écologie pour l'Europe (Brice Lalonde),
LO: lutte ouvrière (Arlette Laguiller),
Maj: union UDF RPR (Dominique Baudis),
MRG: mouvement des radicaux de gauche (Bernard Tapie),
OutM: rassemblement de l'outre mer et des minorités (Ernest Moutoussamy),
PCF: parti communiste français (Francis Wurtz),
PLN: parti de la loi naturelle (Benoît Frappé),
PS: parti socialiste (Michel Rocard),
PT: parti des travailleurs (Daniel Gluckstein),
PVE: politique de vie pour l'Europe (Christian Cotten),
REG: liste régionaliste et fédéraliste (Max Simeoni),
Sara: l'Europe commence à Sarajevo (Léon Schwartzenberg),
Vert: union des écologistes pour l'Europe, les verts (Marie Anne Isler Béguin).



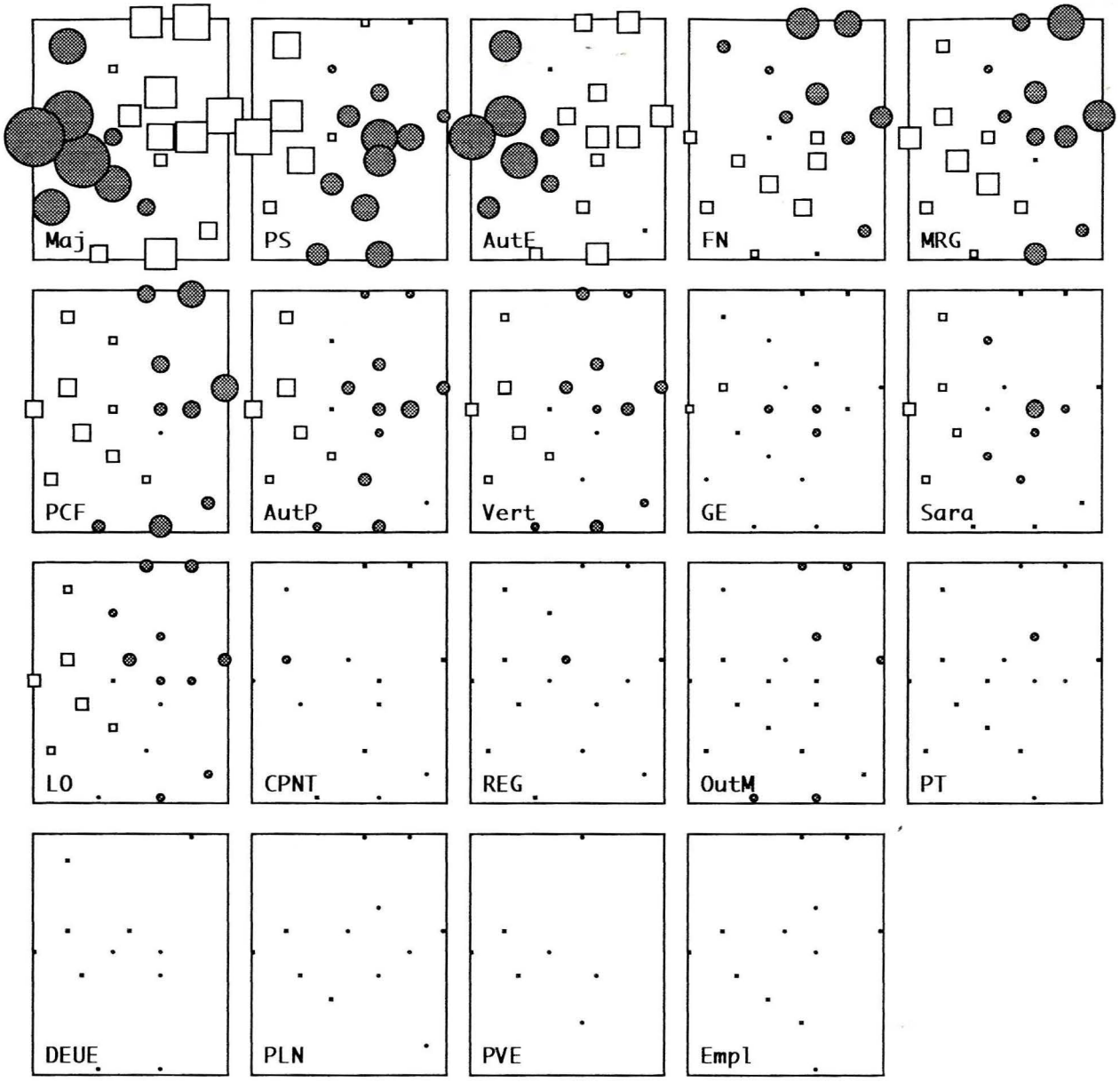
Les arrondissements parisiens et leurs voisinages.

Les données centrées ont été cartographiées dans le graphe n° 1: chacun des rectangles représente Paris, les arrondissements sont matérialisés par un rond (valeur positive) ou un carré (valeur négative), la dimension des ronds et des carrés est proportionnelle à la valeur matérialisée.

Nous avons réalisé le test de Geary sur les 19 variables. Ce test a montré que toutes les variables ont une autocorrélation spatiale évidente.

Les données sont formées:

* d'un graphe de voisinage,
* d'un tableau de variables attributaires simples quantitatives.



Cartographie des listes

Graphe n°1

Nous pouvons donc réaliser une ACP sur le tableau des données attributaires: l'aspect géographique du problème n'est pas pris en compte, ainsi qu'une ACP locale et globale.

b. Réalisation de l'ACP.

Lors de la réalisation de l'ACP, nous avons choisi de ne pas réduire les données. Les variables étant toutes exprimées dans la même unité (en % des votes), cette option était préférable afin de ne pas donner la même importance à toutes les listes, petites ou grandes. Cela permet aussi de minimiser l'impact des erreurs faites sur les petites listes.

Les résultats sont les suivants:

* inertie totale: 88.1 (100%)
* inertie sur le premier axe: 78.5 (89.1%)
* inertie sur le deuxième axe: 8.7 (9.9%)
* inertie sur le troisième axe: 0.38 (0.4%)

Le premier axe (graphe n°2) peut s'interpréter comme une opposition Majorité et Autre Europe face aux autres listes, le deuxième axe comme une opposition PS contre FN et dans une moindre mesure MRG et PCF. La situation MRG FN se retrouve dans les données: ces deux listes semblent avoir le même type d'électorat au niveau de Paris. Le troisième axe (graphe n°3), représentant une inertie faible, oppose le MRG au FN. Les autres listes n'ont que très peu d'importance du fait de la faiblesse de leurs résultats et sont donc mal représentées dans l'ACP sur les données non réduites.

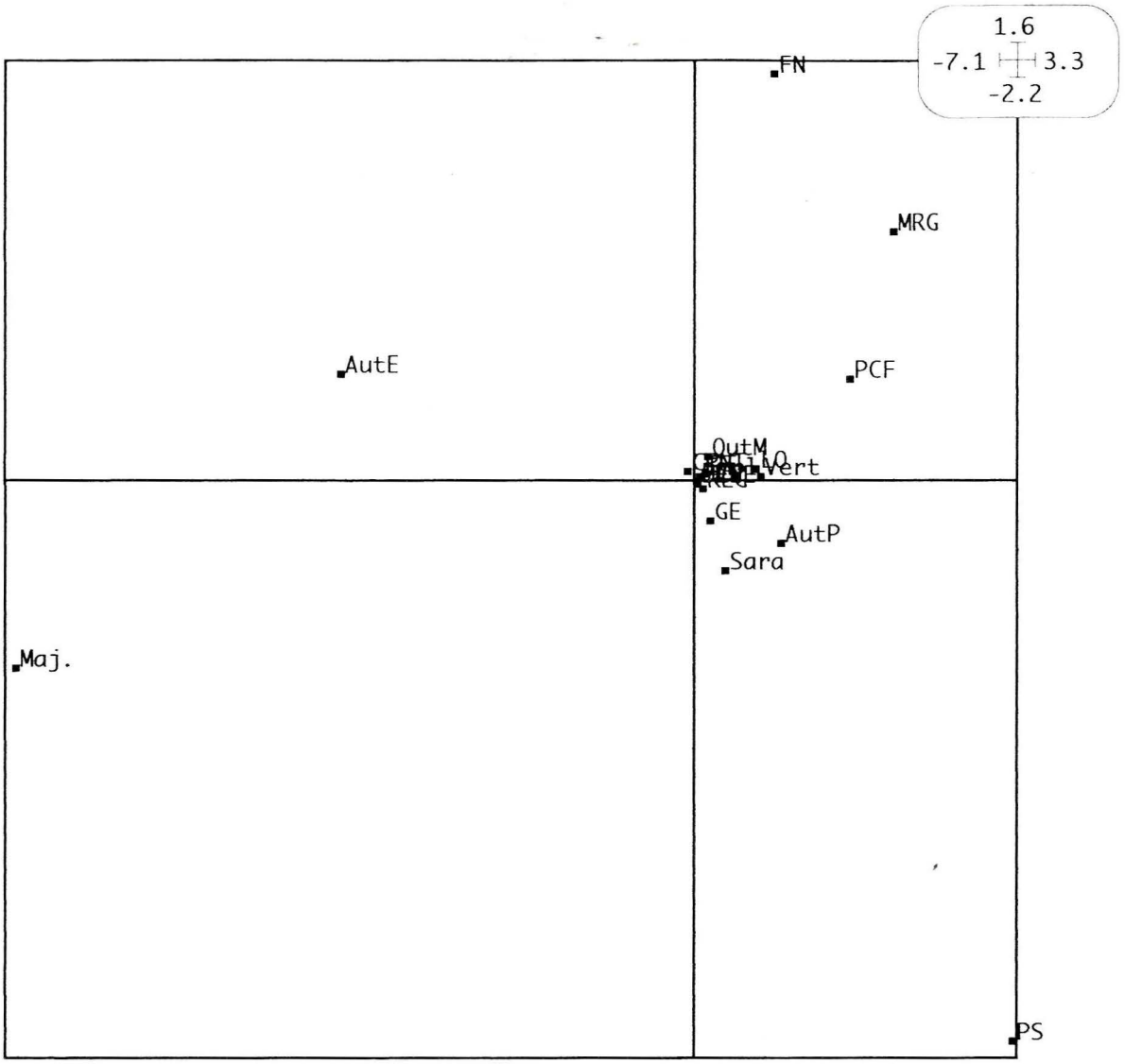
Dans le plan principal de l'espace des individus (graphe n°4), nous avons représenté le graphe de voisinage. D'une façon globale, on retrouve une opposition Est - Ouest: cartiers aisés (16, 17, 8, 7) contre cartiers ouvriers. On note, de plus, une nette déformation du centre de Paris (arrondissements 1,2,3 et 4), cartiers étudiants vers le quadrant en bas à droite (PS). Le plan 2 - 3 (graphe n°5) de l'ACP est difficilement interprétable en dehors des positions du 18ème arrondissement qui a voté FN et du 19ème qui a voté MRG.

c. Réalisation de l'ACP locale.

Dans les légendes des graphes, l'analyse locale est appelée "Geary".
Les résultats de l'analyse locale sont les suivants:

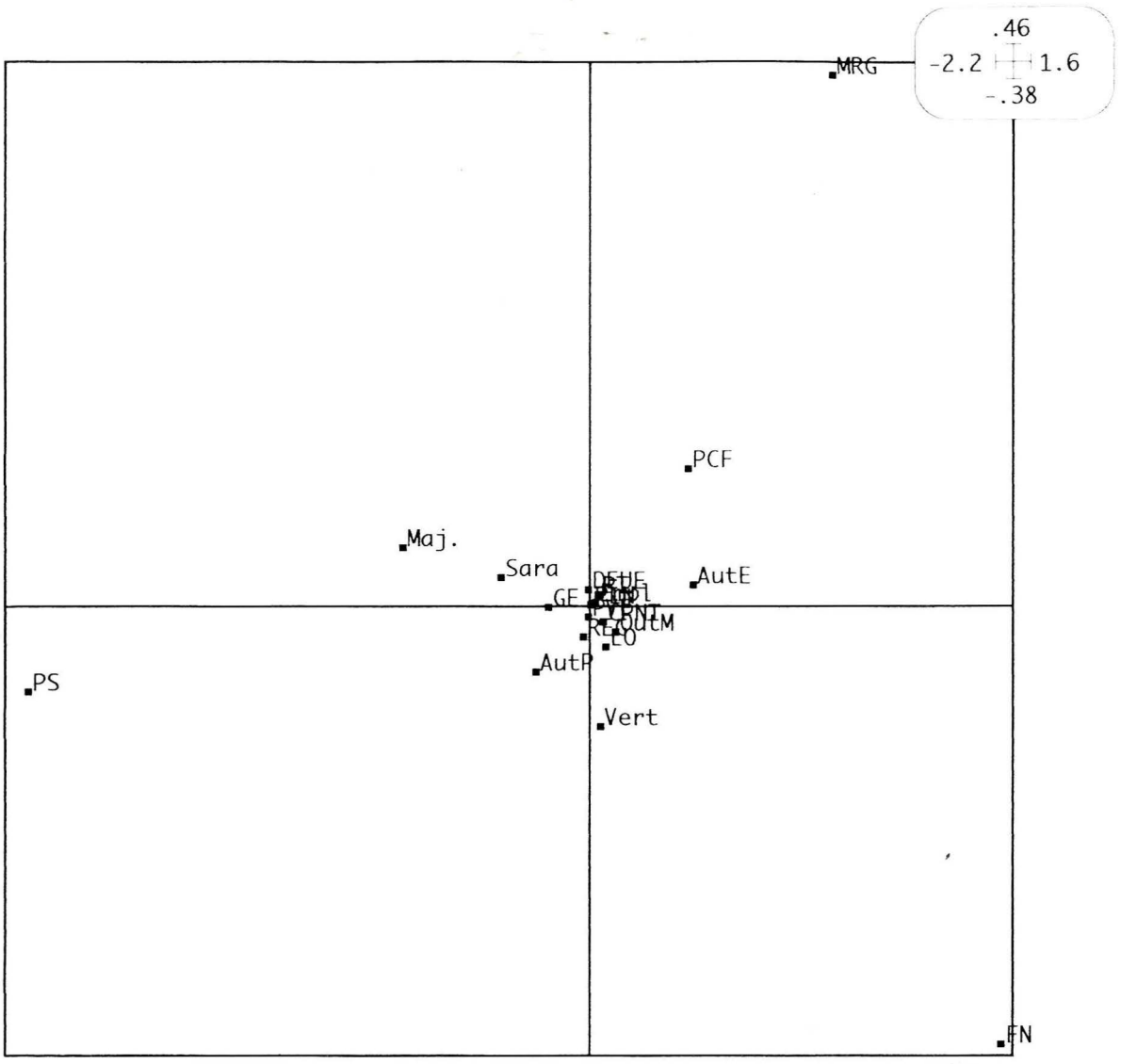
* inertie locale totale: 31.7 (100%)
* inertie sur le premier axe: 26.6 (83.9%)
* inertie sur le deuxième axe: 4.2 (13.3%)
* inertie sur le troisième axe: 0.4 (1.3%)

Les trois axes de l'ACP locale sont les mêmes que ceux de l'ACP simple, seuls les axes deux sont de signes opposés. Ces axes renferment donc la même information, mais l'inertie totale est plus faible.
Nous n'avons pas joint les graphes car ils sont exactement superposables à ceux de l'ACP.



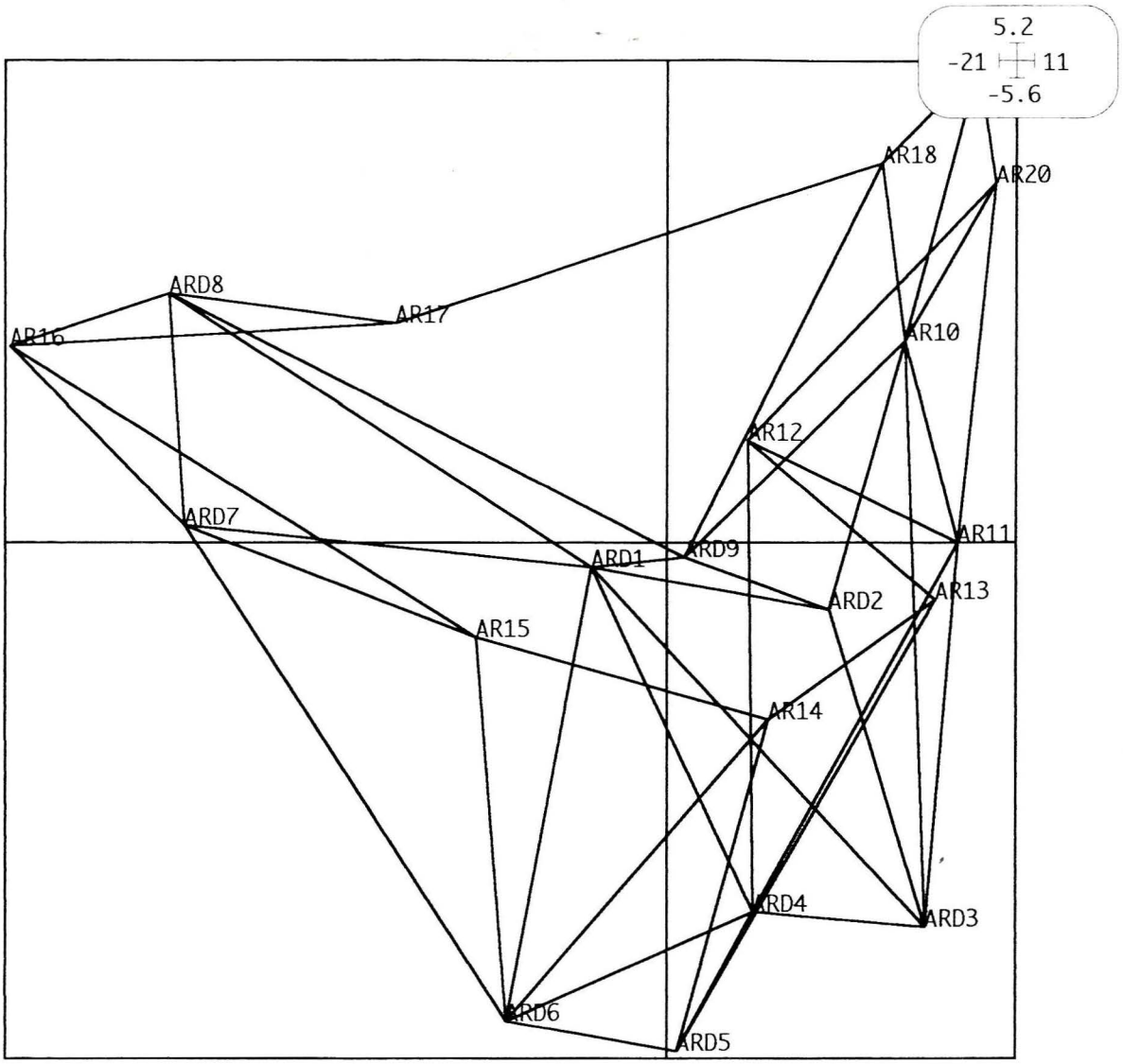
ACP

Graphe n°2



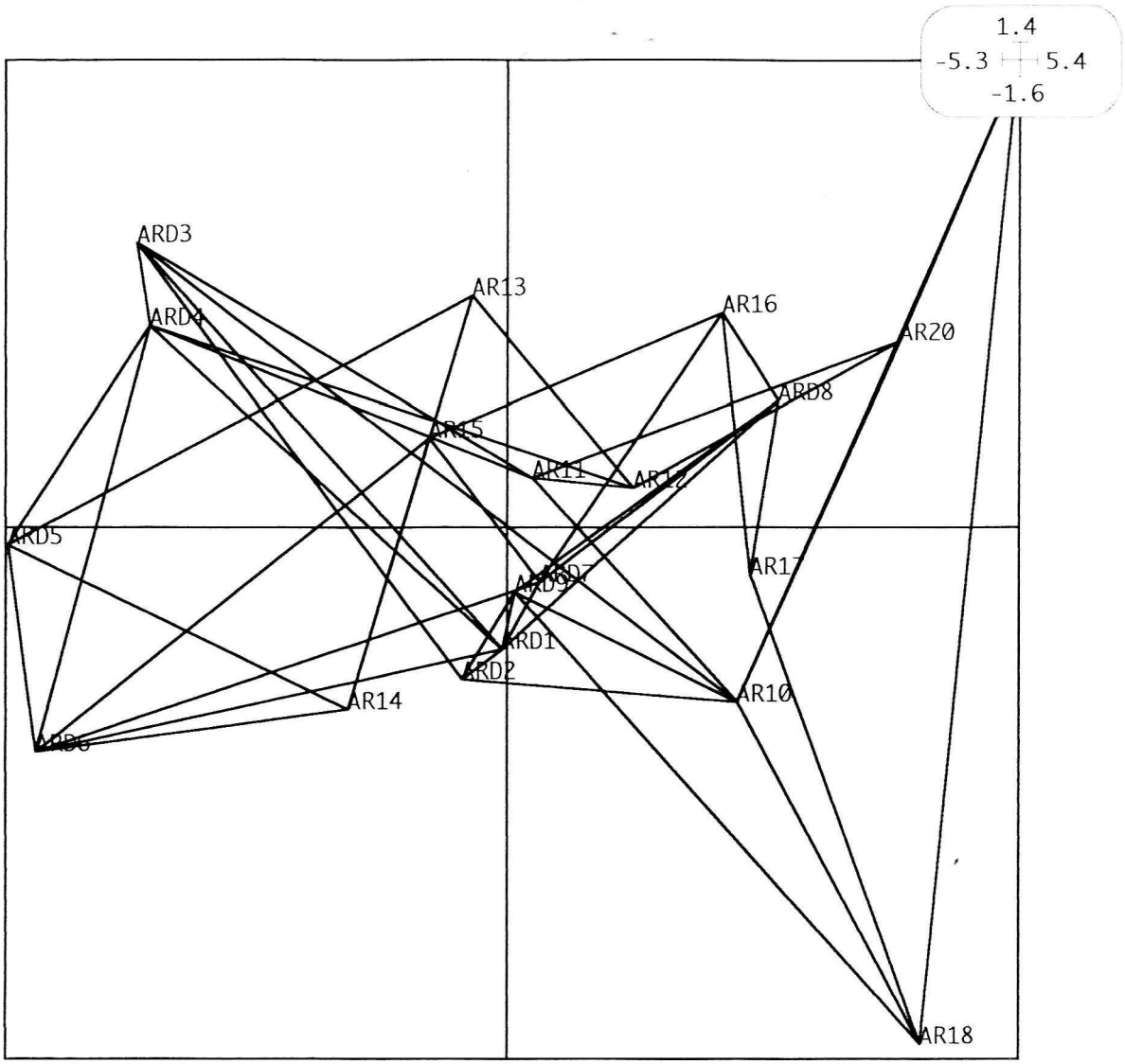
ACP (plan 2-3)

Graphe n°3



ACP

Graphe n°4



ACP (plan 2-3)

Graphe n°5

d. Réalisation de l'ACP globale.

Dans les légendes des graphes, l'analyse globale est appelée "Moran".
Les résultats de l'analyse globale sont les suivants:

* inertie globale totale: 56.5 (100%)
* inertie sur le premier axe: 51.9 (91.9%)
* inertie sur le deuxième axe: 4.5 (8.0%)
* inertie sur le troisième axe: 0.07 (0.13%)

Comme dans le cas précédant, les axes 1 et 2 de l'ACP globale sont les mêmes que ceux de l'ACP. Par contre le troisième axe est différent de celui de l'ACP (graphe n°6): il oppose le PCF à la liste Sara et GE. Ce sont trois petites listes, il en résulte que l'inertie sur cet axe est très faible. La représentation des individus dans le plan 2-3 (graphe n° 7) est très différente de celle de l'ACP: les voisinages semblent plus respectés.

e. Analyse des résultats.

Sur cet exemple, les trois analyses donnent les mêmes résultats pour la majorité de l'inertie. Sur le plan principal commun, certains individus proches géographiquement sont éloignés (exemple: les arrondissement 6 et 7) et d'autres éloignés géographiquement sont rapprochés (exemple: les arrondissements 2 et 14). La contrainte géographique imposée par le graphe de voisinage ne semble pas dominante.

Pour juger les qualités des axes nous les avons cartographiés (graphe n°8). Les deux premiers axes des trois analyses sont identiques et ont de bonnes qualités de cartographie: ils délimitent des zones géographiques homogènes avec un passage progressif des valeurs positives aux valeurs négatives.
Pour les troisièmes axes de ces analyses, seul l'analyse globale garde des qualités de cartographie et prend un avantage sur l'ACP. Cet avantage n'est toutefois pas important en raison de la faiblesse de l'inertie représentée par le troisième axe.

L'équivalence des trois méthodes peut s'expliquer par le fait que les données initiales étaient, sans aucune modification, directement cartographiables. Ce résultat se retrouve en étudiant le graphe n° 1: les résultats des grandes listes définissent des zones homogènes.

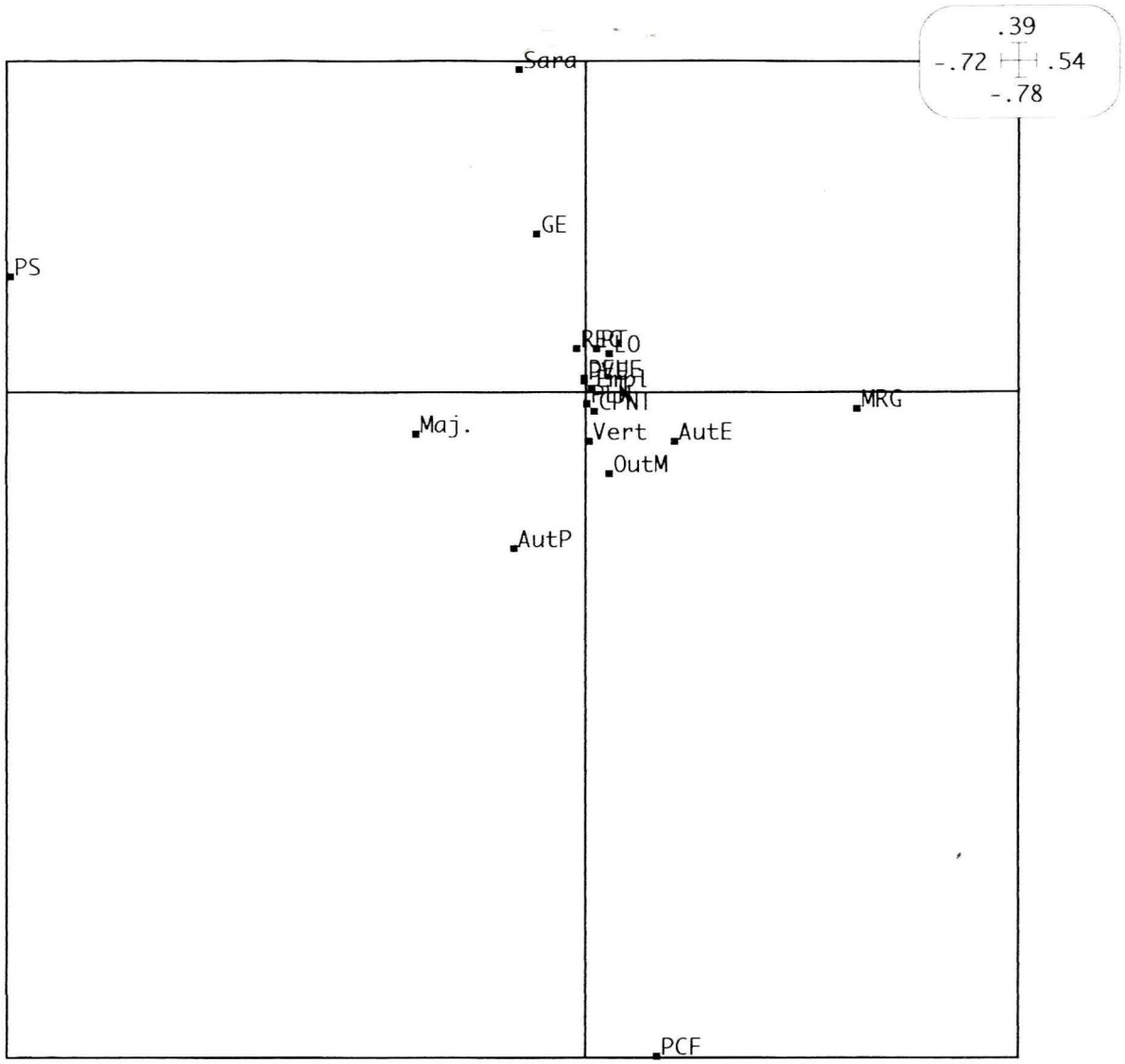
2. Deuxième exemple: simulation sur Paris.

a. Présentation des données.

Dans cet exemple, nous avons repris le support géographique des arrondissements de Paris.

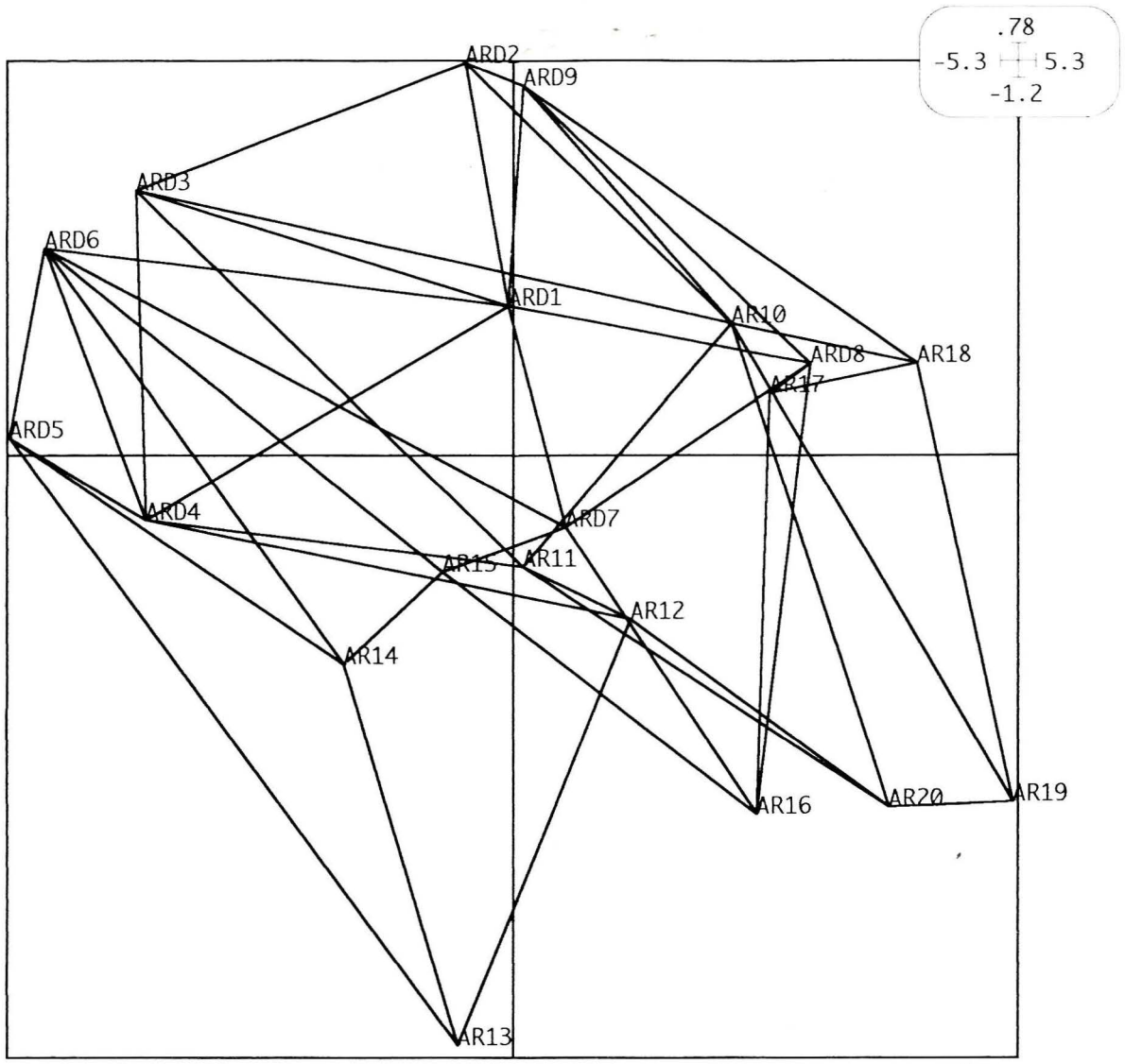
Nous avons créés 5 variables aléatoires suivant une loi normale, et deux variables n° 6 et 7 avec un bruit aléatoire et une tendance géographique. La tendance était:

* Nord - Sud pour la variable 6,
* Ouest - Est pour la variable 7.



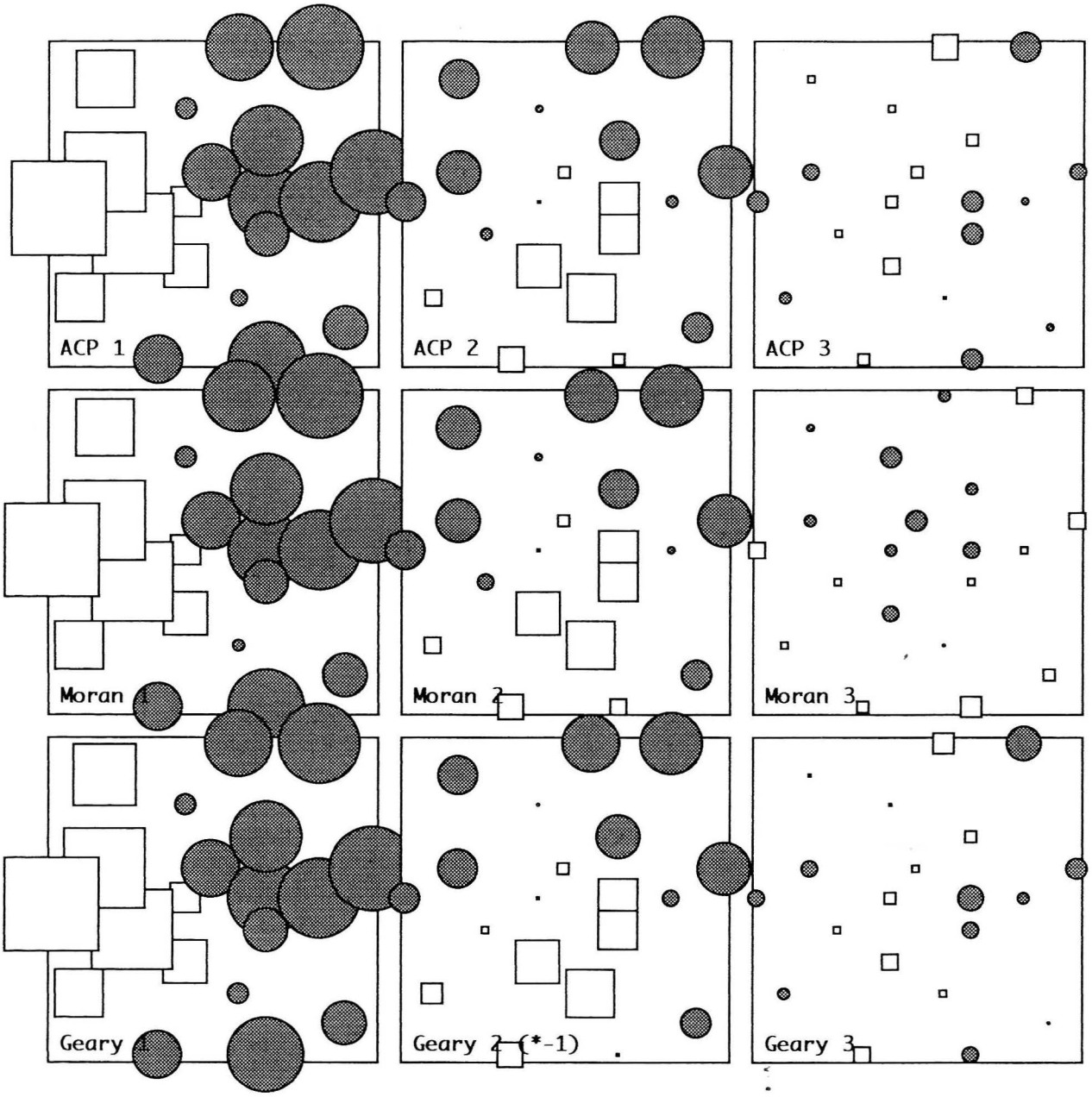
Moran (plan 2-3)

Graphe n°6



Moran (plan 2-3)

Graphe n°7



Cartographie des composantes

Graphe n°8

Il faut noter qu'en raison du bruit aléatoire, ces tendances ont été déformées et atténuées. Les données brutes sont dans le tableau n° 2, nous avons cartographié ces variables (graphe n° 9) pour une meilleure compréhension.

b. Mise en oeuvre des analyses.

Nous avons centré et réduit les données avant de réaliser les trois analyses afin de donner la même importance à toutes les variables.

	Inertie totale	1er axe	2ème axe	3ème axe
ACP	7	26.0 %	21.1 %	17.7 %
ACP locale	6.68	29.2 %	22.8 %	21.1 %
ACP globale	0.32	68.3 %	17.5 %	14.3 %

On note que l'inertie locale est forte, ce qui correspond au fait que, les variables étant aléatoires, les voisins n'ont qu'une très faible probabilité de se ressembler. L'inertie globale n'est que de 0.32, mais elle ne pouvait pas être supérieure à 2 du fait qu'il n'y ait que deux variables cartographiables.

Cette fois, les plans principaux sont très différents (graphes 10 à 15), mais sont difficilement interprétables du fait de l'absence de signification des variables 1 à 5. Pour l'ACP les variables 6 et 7 définissent l'axe 2. pour l'analyse locale les variables 6 et 7 s'opposent sur l'axe 2. Pour l'analyse globale, ces deux variables sont fortement corrélées avec l'axe 1 et, dans une moindre mesure, avec l'axe 2 pour la variable 7. Il semblerait donc que la majorité de l'inertie globale provienne bien des deux dernières variables.

c. Analyse des résultats.

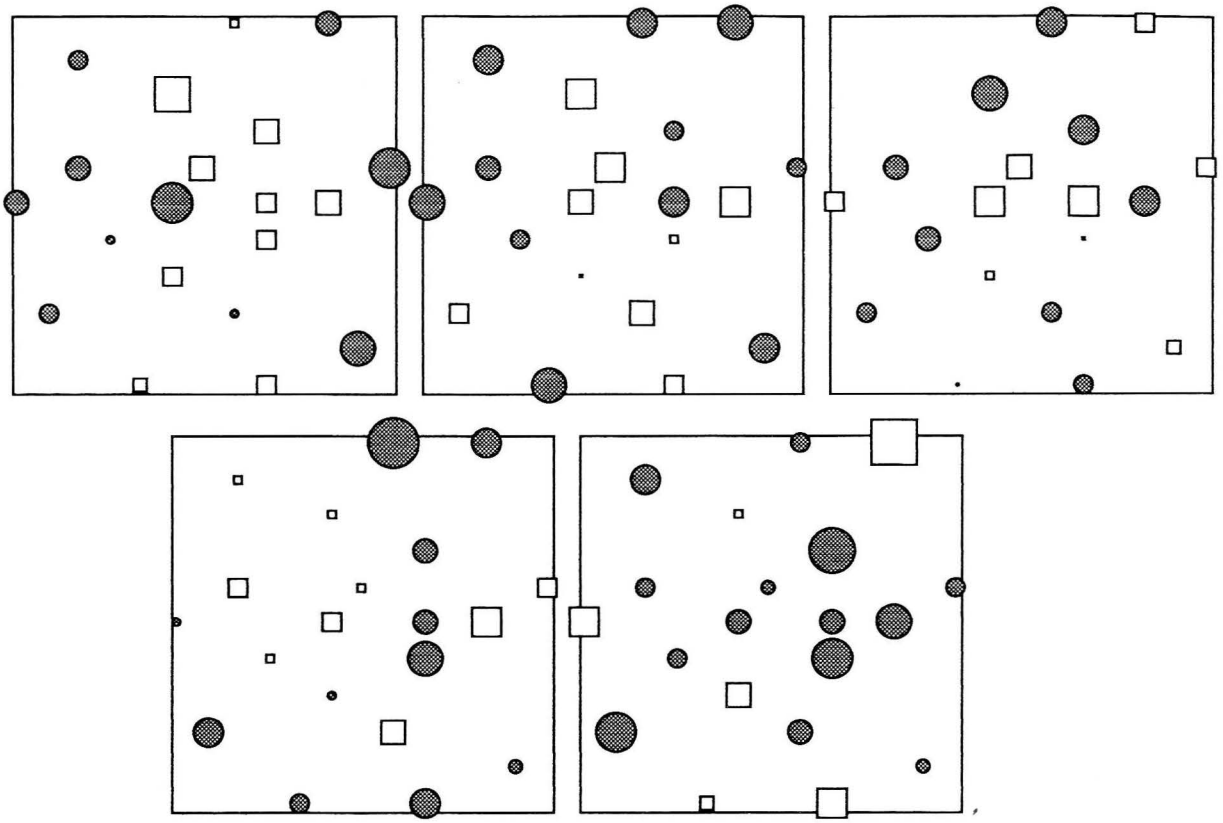
Nous avons cartographié les composantes des trois analyses (graphe n° 16). Aucune des composantes de l'ACP ou de l'analyse locale ne définit de zones géographiques homogènes, alors que les deux premières de l'analyse globale le font. La première composante de l'analyse globale défini une opposition Nord-Ouest Sud-Est, et la deuxième une zone centrale décalée vers le Nord-Ouest qui s'oppose à sa périphérie. On ne retrouve donc pas exactement nos deux variables 6 et 7, leurs corrélations avec les axes de l'analyse globale le montraient, mais cela peut être dû aux variables initiales qui n'étaient pas exactement orthogonales. L'analyse globale trouve deux composantes cartographiables seulement.

Le plan principal de l'analyse locale permet de mettre en évidence les voisins qui sont les plus différents entre eux. Par exemple, sur le graphe n° 13 on voit que les arrondissements 19 et 18, 18 et 9, bien que voisins, sont très différents. L'individu n° 1 semble avoir une position particulière puisque qu'il s'éloigne de la même manière de tous

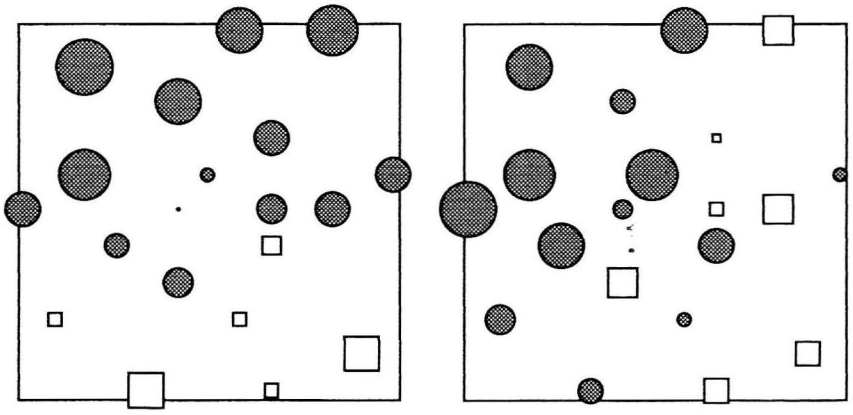
Données aléatoires sur les arrondissements de Paris

Arrondissements	Var. 1	Var. 2	Var. 3	Var. 4	Var. 5	Var. 6	Var. 7
1	1.93	-1.07	-1.82	-0.76	0.65	0.51	0.05
2	-0.98	-1.54	-1.16	-0.13	0.30	2.80	0.29
3	-0.75	0.90	-1.78	0.59	0.71	-0.46	0.97
4	-0.85	-0.15	-0.03	1.55	1.62	1.35	-0.71
5	0.15	-1.03	0.47	-1.14	0.70	0.34	-0.36
6	-0.74	-0.08	-0.19	0.11	-1.20	-1.60	1.18
7	0.10	0.56	0.66	-0.24	0.37	2.55	0.78
8	0.59	0.70	0.78	-0.83	0.55	2.98	2.60
9	-2.13	-1.50	1.61	-0.12	-0.29	0.85	2.12
10	-1.42	0.57	0.90	0.62	2.25	-0.27	1.24
11	-1.22	-1.45	0.87	-1.44	1.39	-1.75	1.28
12	1.50	0.96	-0.36	0.20	0.27	-1.09	-2.13
13	-0.77	-0.95	0.55	0.89	-1.47	-1.02	-0.31
14	-0.52	1.39	0.02	0.39	-0.48	0.81	-2.15
15	0.55	-0.81	0.46	0.96	1.84	1.13	-0.55
16	0.59	1.31	-0.85	0.15	-1.63	3.28	1.42
17	0.52	0.89	-0.01	-0.18	1.16	2.56	3.55
18	-0.21	0.90	1.02	3.11	0.52	2.14	2.41
19	0.61	1.58	-0.77	0.93	-3.45	-1.59	2.83
20	1.74	0.54	-0.73	-0.95	0.55	0.18	1.42

Tableau n° 2

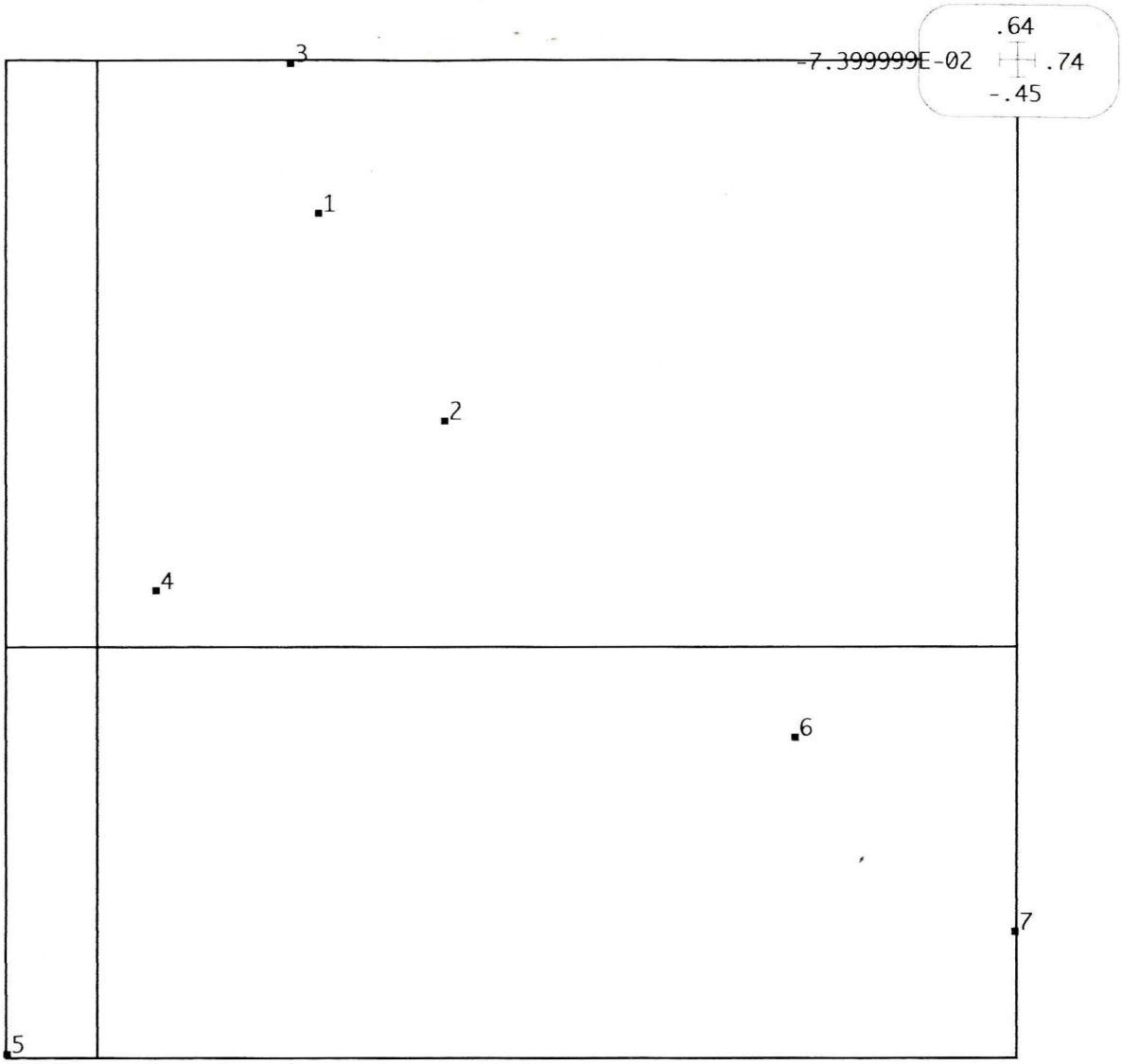


5 variables aléatoires



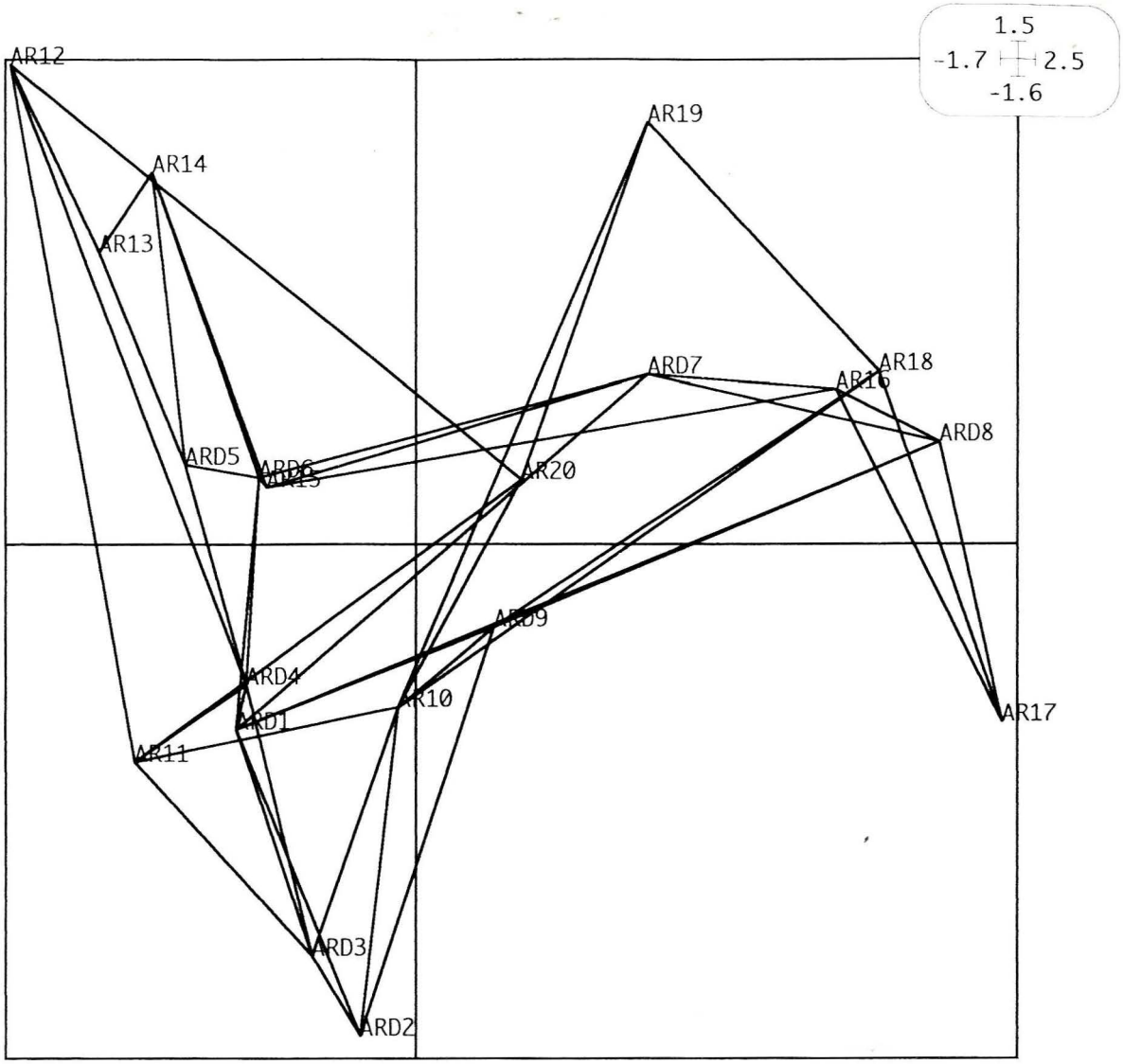
2 variables avec tendance géographique
et bruit aléatoire

Graphe n°9



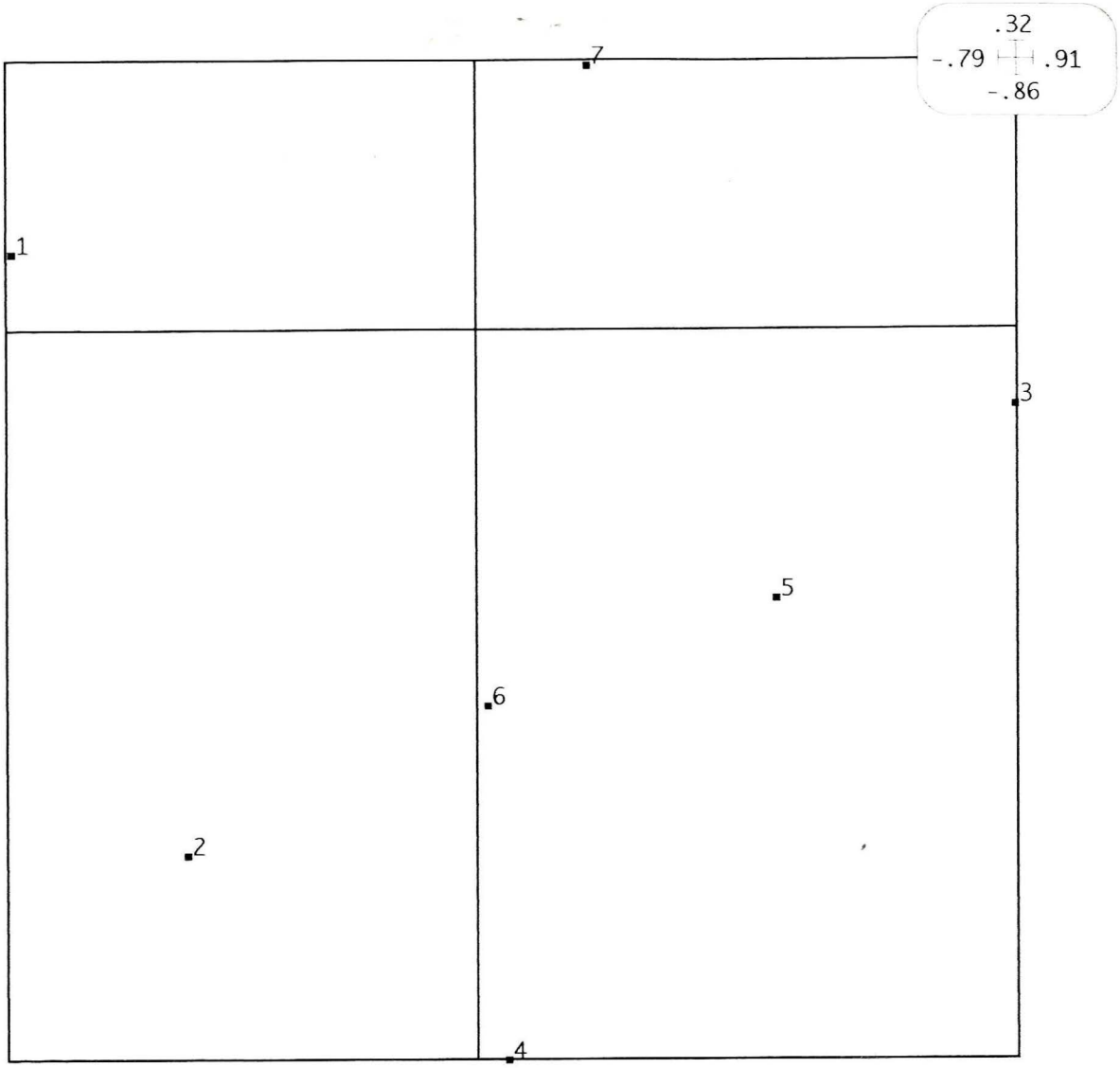
Moran (plan 1-2)

Graphe n°10



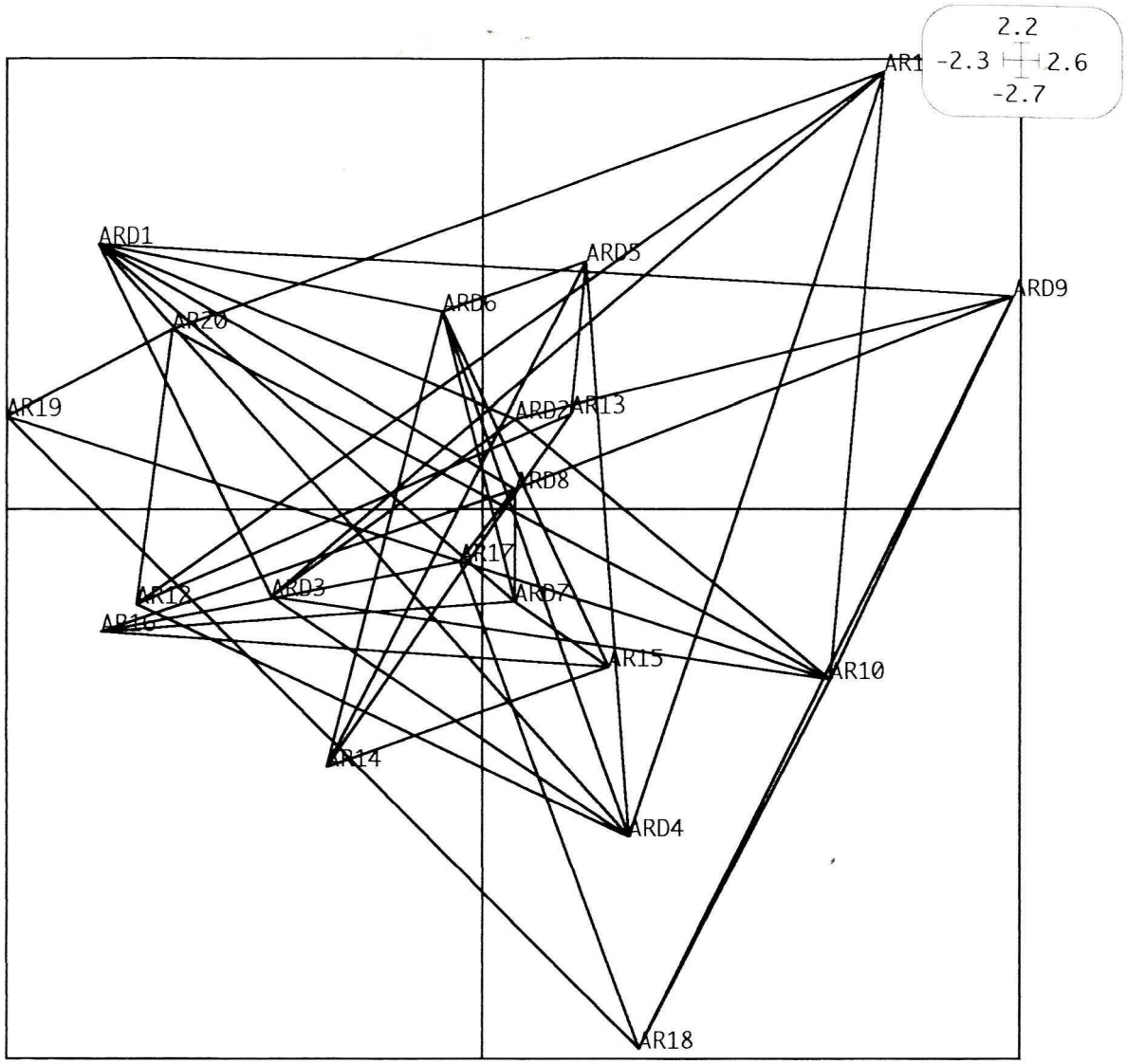
Moran (plan 1-2)

Graphe n°11



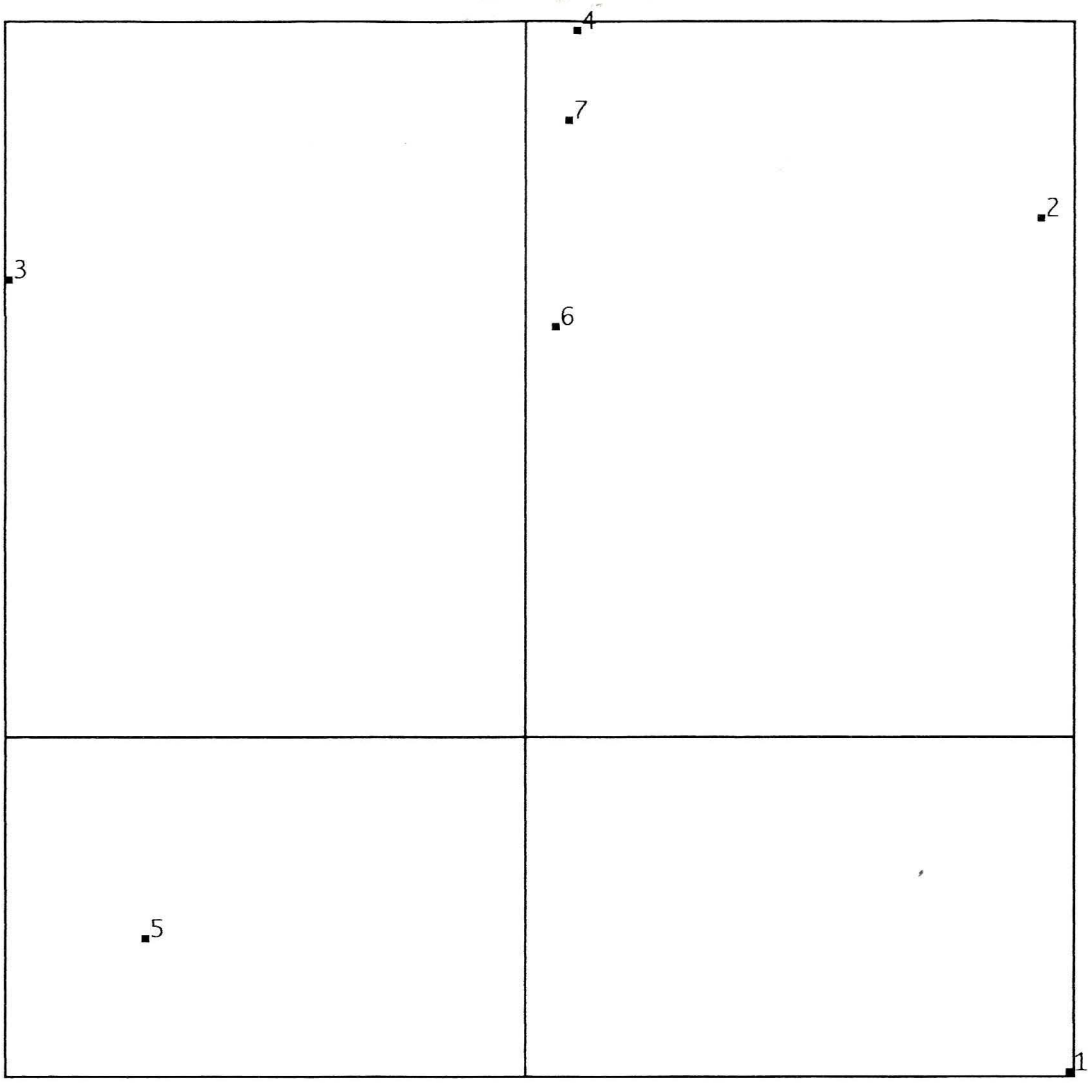
Geary (plan 1-2)

Graphe n°12



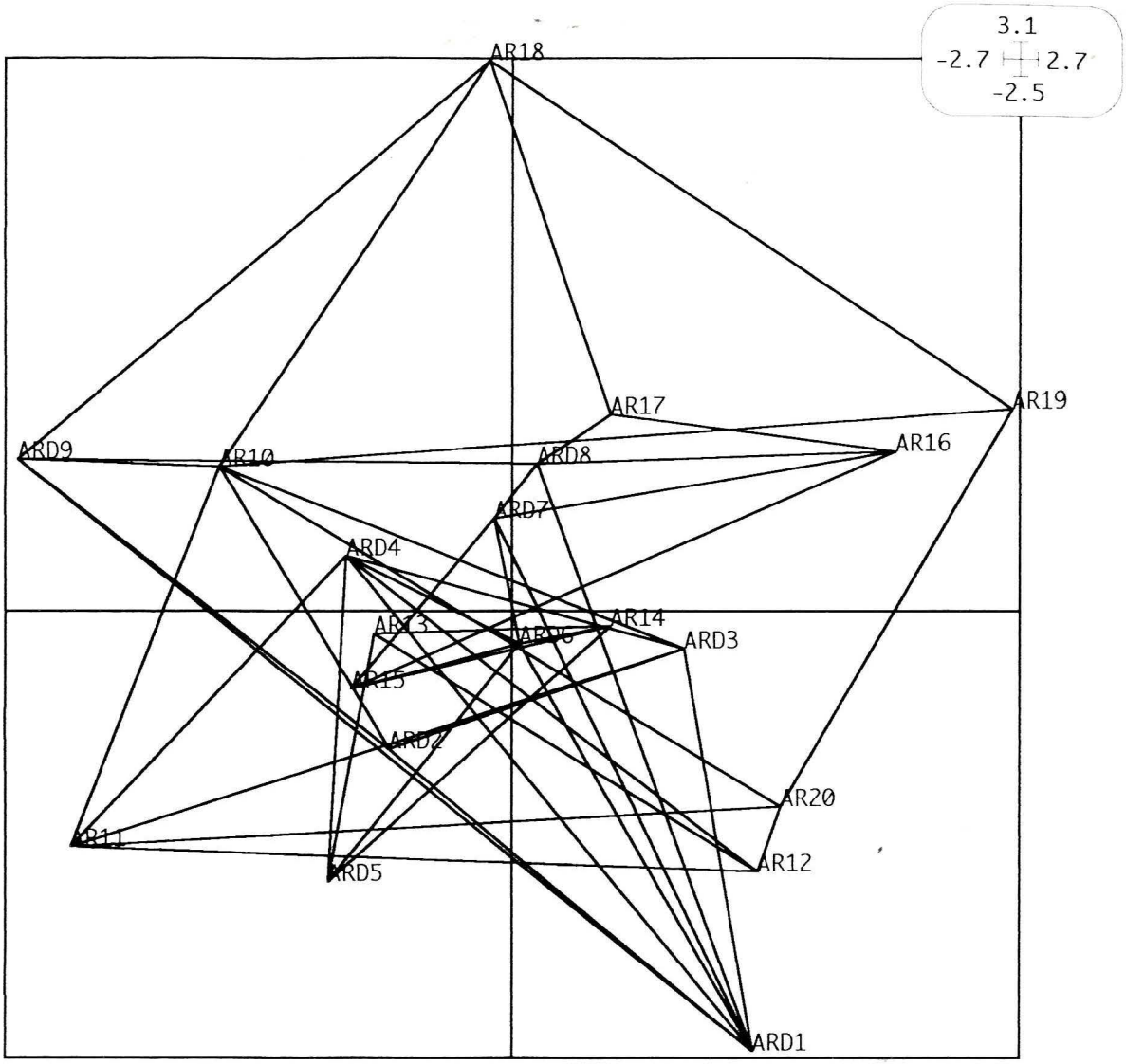
Geary (plan 1-2)

Graphe n°13



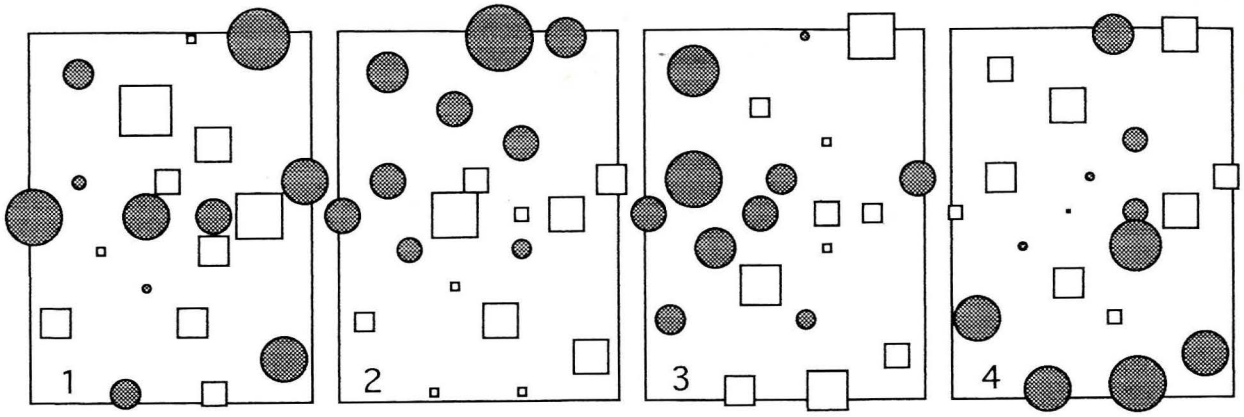
ACP (plan 1-2)

Graphe n°14

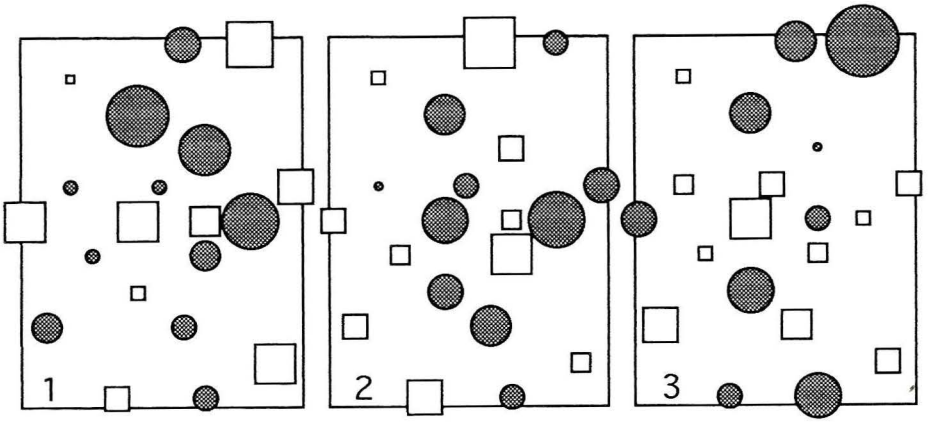


ACP (plan 1-2)

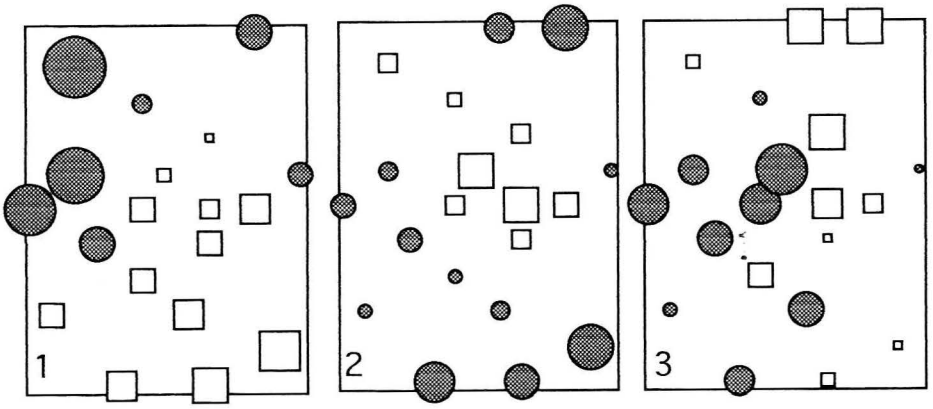
Graphe n°15



ACP composantes 1,2,3 et 4



Geary composantes 1,2 et 3



Moran composantes 1,2 et 3

Graphe n°16

ses voisins. Nous ne pouvons expliquer cette situation puisque les variables initiales n'ont pas de signification.

Ce dernier exemple montre donc que l'analyse globale est plus efficace dans la recherche de composantes cartographiables que l'ACP simple. Elle peut donc être utilisée dans les SIG avec cet objectif. L'analyse locale est utile à la recherche de voisins très différents entre eux.

B. Exemple d'utilisation de l'ACPVI.

1. Données utilisées.

Nous avons repris le jeux de données des élections européennes. Ces données ont été, pour les mêmes raisons que précédemment, centrées et non réduites. Ces données forment le tableau à expliquer.

Les variables explicatives étaient formées par les coordonnées X et Y des arrondissements de Paris. Nous avons complété ce tableau de variables explicatives par deux variables supplémentaires X^2 et Y^2, cela afin d'obtenir des combinaisons linéaires du second degrés par rapport à X et Y.

2. Résultats obtenus.

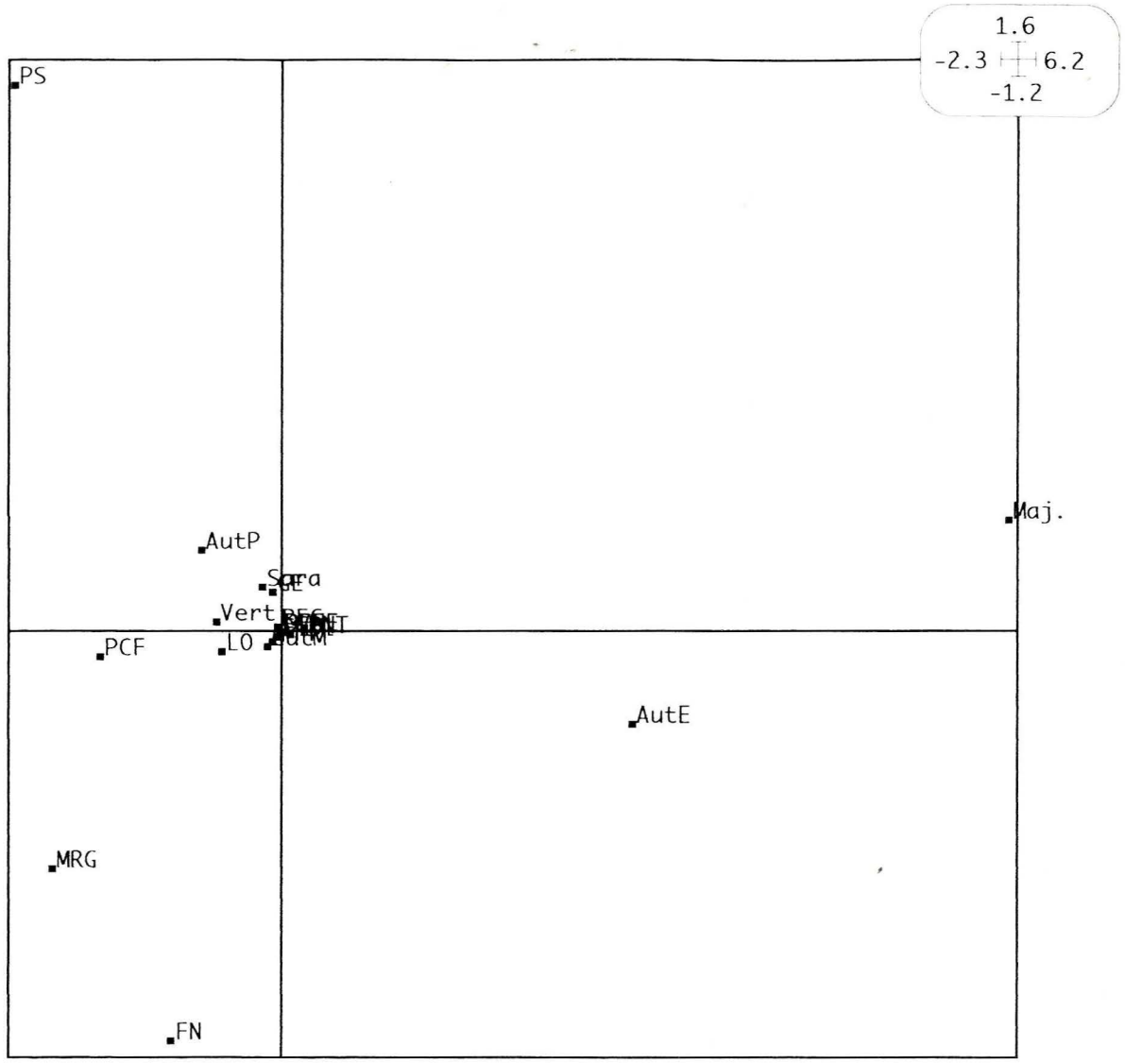
L'inertie totale du nuage des points est de 88.1, et l'inertie expliquée par l'ACPVI est de 63.8. On a donc expliqué 72.4 % de l'inertie totale, ce qui est énorme compte tenu de la simplicité de nos variables explicatives.

Le plan principal de l'ACPVI (graphe n° 17) est très proche de celui de l'ACP, il représente 99.8 % de l'inertie expliquée. La représentation des individus (graphe n° 18) restitue parfaitement la géographie de Paris, on y a fait figurer, à titre indicatif, le graphe de voisinage. Cela démontre que les composantes de l'ACP étaient facilement cartographiables. La représentation des composantes de l'ACPVI (graphe n° 19) montre que le premier axe est très corrélé avec X, et le deuxième avec Y. En conséquence, les composantes de l'ACPVI sont parfaitement cartographiables.

3. Utilisations de l'ACPVI dans les SIG.

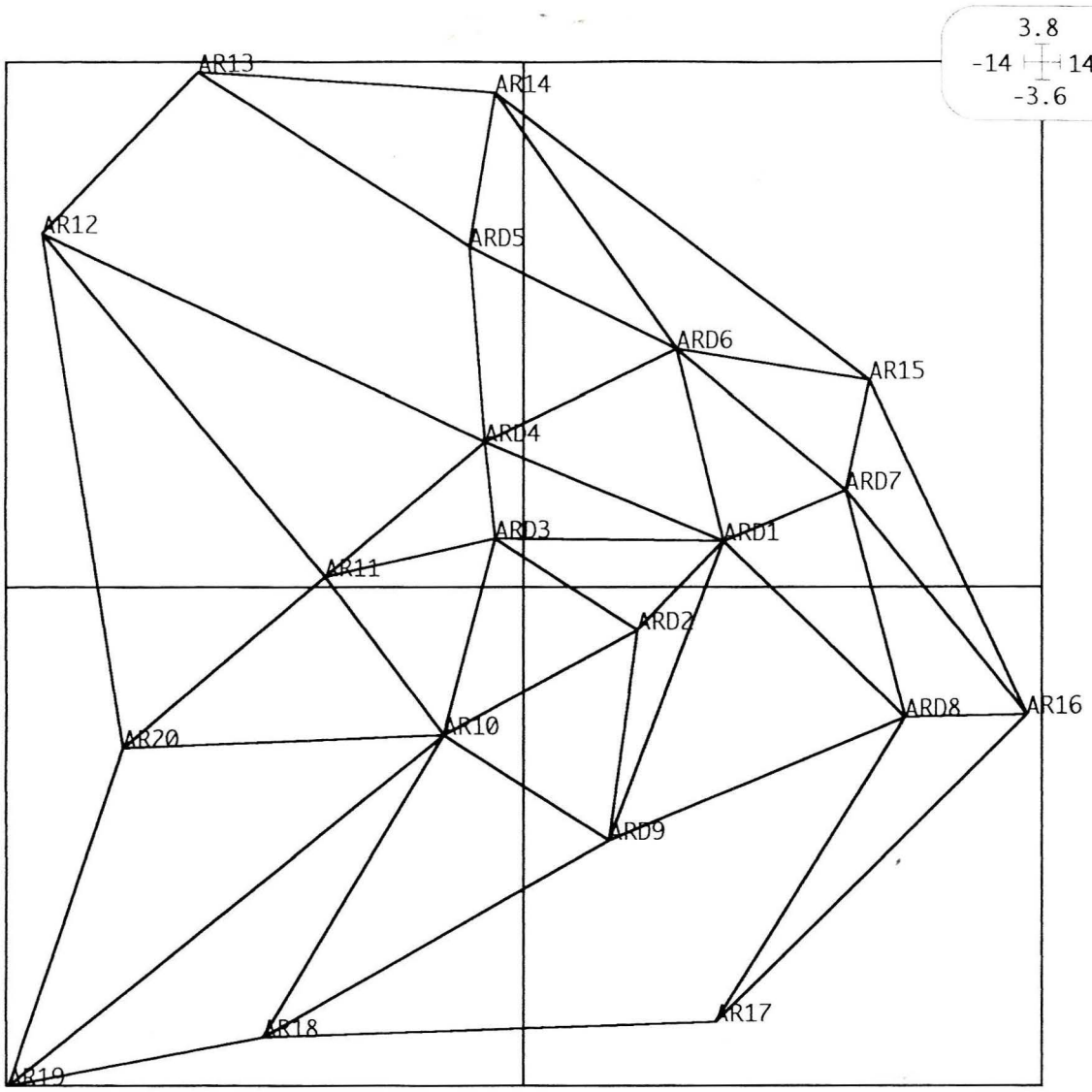
Telle que nous l'avons utilisée, l'ACPVI ne peut donner de bons résultats que si les données brutes sont directement cartographiables. Dans toute autre situation, le risque est de ne pas expliquer suffisamment d'inertie. Dans tous les cas, les composantes obtenues sont cartographiables car elles sont combinaisons linéaires de X, Y, X^2 et Y^2, mais elles peuvent ne rien représenter en terme d'inertie. L'exemple des élections est un cas particulier. A la suite de cette utilisation, il est possible de reconstituer les données, ce qui revient à faire une modélisation (graphe n° 20). On sort, avec cette démarche, du cadre de l'objectif exploratoire que nous nous étions fixé.

Une utilisation plus rationnelle de l'ACPVI peut être faite dans le cadre des SIG. Il s'agit de construire un tableau de données explicatives plus complexe. Dans l'exemple des



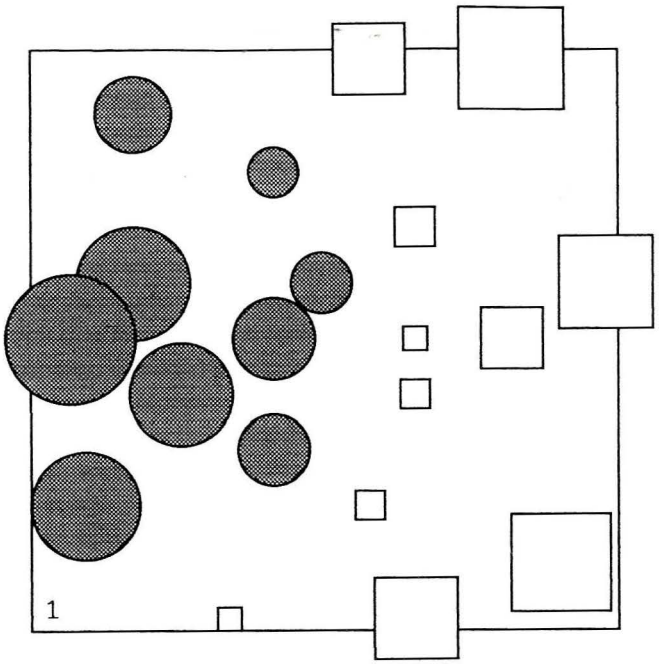
ACPVI (plan 1-2)

Graphe n°17

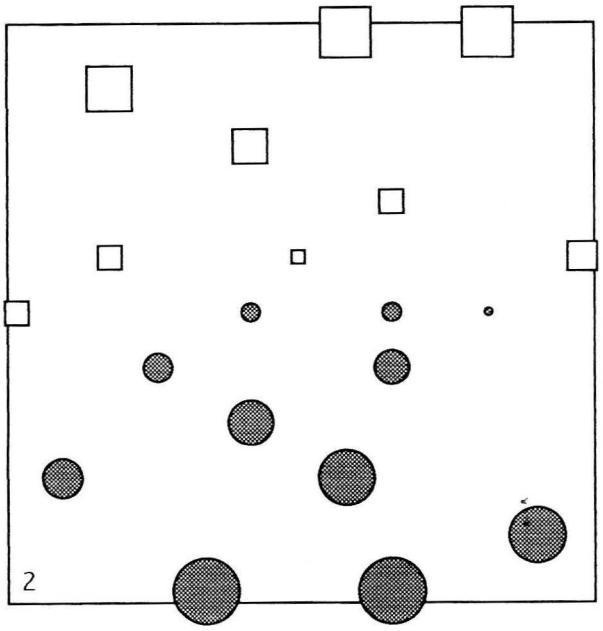


ACPVI (plan 1-2)

Graphe n°18

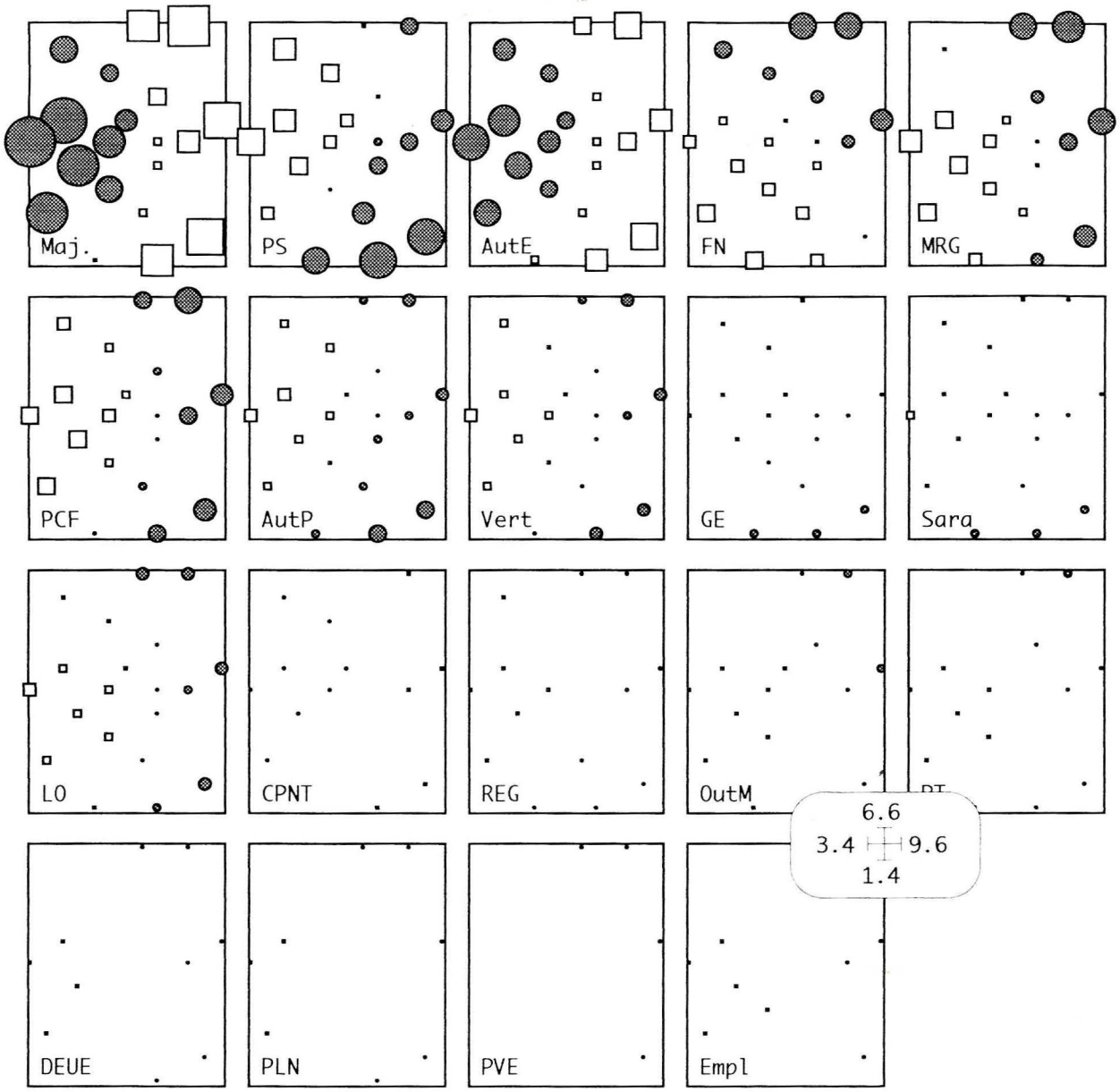


Composante 1 de l'ACPVI



Composante 2 de l'acpvi

Graphe n°19



Reconstitution des données après ACPVI

Graphe n°20

élections européennes, nous aurions pu utiliser des variables explicatives décrivant la population de Paris par arrondissement. Ce sont ces variables qui, du fait de leurs forts liens avec la géographie, donnent la dimension géographique de l'analyse.

Conclusion

Cette étude démontre qu'il est possible de développer, dans la cadre des Systèmes d'Information Géographique, des méthodes statistiques complexes. En effet, les outils de requête sur base de données, proposés dans les SIG, sont capables de construire des tableaux statistiques volumineux qu'il est nécessaire d'exploiter par des méthodes analytiques et synthétiques sans perdre les caractéristiques géographiques.

Toutefois, il s'avère, à la lumière de la première partie, que l'interface méthodes - données géographiques et attributaires est le point le plus délicat de toute étude. Cette interface (choix des individus statistiques et données géographiques) doit être faite au cas par cas. C'est la limite au développement de modules informatiques spécifiques aux SIG.

Dans cette étude, nous avons exposé un certain nombre de méthodes, ayant un objectif exploratoire, dérivées de l'analyse des données. Il faut noter qu'il existe de nombreuses autres méthodes notamment dans le but de modéliser ou de prévoir des données.

Dans tous les cas, l'interprétation des résultats de ces analyses reste délicate. Une mauvaise interprétation peut amener facilement à des conclusions aberrantes. Cette étape doit absolument être menée par une personne sensibilisée aux problèmes statistiques.

Bibliographie

[1] Benali H. et Escofier B. (1989). Smooth factorial analysis and factorial analysis of local differences. In *Multiway Data Analysis*. R. Coppi and S. Bolasco editors. Elsevier Science Publishers B.V. (North Holland) : 327 - 339.

[2] Cailliez F. et Pages J.P. (1976). Introduction à l'analyse des données. Edt. Smash : 1 - 616.

[3] Carlier A. (1985). Application de l'analyse factorielle des évolutions et de l'analyse intra - périodes. *Statistiques et Analyse de données* : 10, 27 - 53.

[4] Carlier A. (1985). Analyse des évolutions sur table de contingence : quelques aspects opérationnels. Quatrièmes journées internationales analyse de données et informatique. Cahier INRIA, tome 2, 421 - 428.

[5] Chessel D. et Mercier P. (1993). Couplage de triplets statistiques et liaisons espèces environnement. *Biométrie et Environnement*. Lebreton J.D. et Asselein B. éditeurs. 15 - 43.

[6] Chessel D. et Sabatier R. (1994). Couplage de triplets statistiques et graphes de voisinage. En cours de parution. edt Masson.

[7] Cliff A.D. et Ord J.K. (1973). Spatial autocorrelation. Edt. Chorley & Harvey.

[8] Cornillon P.A. (1992). Analyse spatio - temporelle dans les tableaux à trois dimensions. Rapport de DEA de Biostatistique. ENSA.M Unité de Biométrie, 2 Place Pierre Viala, 34060 Montpellier cédex 1.

[9] Escofier B. et Pagés J. (1990). Analyses factorielles simples et multiples : objectifs, méthodes et interprétation. Edt. Dunod : 1 - 267.

[10] Faray A. (1993). Analyse de contiguïté : une analyse discriminante généralisée à plusieurs variables qualitatives. *Rev. Statistique Appliquée* : 3, 73 - 84.

[11] Fichet B. (1987). The role played by L_1 in data analysis. In Statistical data anlysis based on the L_1 norm and related methods. Y. Dodce editor. Elsevier Science Publishers B.V. (North Holland) : 185 - 193.

[12] Jayet H. (1993). Analyse spatiale quantitative : une introduction. Edt Economica : 1 - 202.

[13] Lavit Ch. (1988). Analyse conjointe de tableaux quantitatifs. Edt. Masson, Méthode et Programmes : 1 - 251.

[14] Lavit Ch. (1988). Présentation de la méthode STATIS permettant l'analyse conjointe de

plusieurs tableaux de données quantitatives. *Les cahiers de la recherche développement* : 18, 49 - 60.

[15] Le foll Y. (1982). Pondération des distances en analyse factorielle. *Statistique et Analyse des données* : 1, 13 - 31.

[16] Méot A., Chessel D. et Sabatier R. (1993). Opérateurs de voisinage et analyse des données spatio - temporelles. *Biométrie et Environnement*. Lebreton J.D. et Asselein B. éditeurs. 45 - 71.

[17] Sabatier R., Lebreton J.D. et Chessel D. (1989). Principal component analysis with instrumental variables as tool for modelling composition data. In *Multiway Data Analysis*. R. Coppi and S. Bolasco editors. Elsevier Science Publishers B.V. (North Holland) : 341 - 352.

[18] Saporta G. (1990). Probabilités, analyse des données et statistique. Edt.Technip : 1 - 493.

[19] Wartenberg D. (1985). Canonical trend surface analysis : a method for describing geographic patterns. *Systematic Zoology* : 34(3), 259 - 279.

[20] Wartenberg D. (1985). Saptial Autocorrelation as Criterion for Retaining Factors in Ordinations of Geographic Data. *Mathematical Geology* : 17, 665 - 682.

[21] Wartenberg D. (1985). Multivariate Spatial Correlation : A Method for Exploratory Geographical Analysis. *Geographical Analysis* : 17, 263 - 283.

RESUME

Les outils de requête sur base de données, proposés dans les Systèmes d'information géographique, permettent de construire des tableaux de données complexes qu'il est nécessaire d'exploiter avec des méthodes statistiques adaptées.

Cette étude cherche à poser les problèmes particuliers liés à la gestion et à l'analyse statistique des données attributaires et géographiques dans les SIG.

Les principes mathématiques de quelques méthodes exploratoires, dérivées de l'analyse des données, ont été développés. Le choix des méthodes statistiques a été dicté par la nécessité de prendre en compte l'aspect géographique du support des données. Une comparaison de trois méthodes à travers leurs résultats sur un exemple précis conclut cette étude.

MOTS CLES

Systèmes d'Information Géographiques (SIG) - Statistiques
Analyse des données - ACP - Voisinage - Analyse locale - Analyse globale- ACPVI - STATIS